(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
12 April 2001 (12.04.2001)

PCT

(10) International Publication Number
WO 01/24820 A1

(51) International Patent Classification⁷: A61K 39/095, 39/02, 39/00, C07K 1/00

(21) International Application Number: PCT/US00/28095

(22) International Filing Date: 10 October 2000 (10.10.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/158,338 7 October 1999 (07.10.1999) US
60/158,425 7 October 1999 (07.10.1999) US

(71) Applicant (*for all designated States except US*): **CORIXA CORPORATION** [US/US]; Suite 200, 1124 Columbia Street, Seattle, WA 98104 (US).

(72) Inventors; and

(75) Inventors/Applicants (*for US only*): **SKEIKY, Yasir** [CA/US]; 8327 25th Avenue, N.W., Seattle, WA 98107 (US). **REED, Steven** [US/US]; 2843 122nd Place, NE, Bellevue, WA 98005 (US). **HOUGHTON, Raymond, L.** [US/US]; 2636 242nd Place, SE, Bothell, WA 98021 (US). **MCNEILL, Patricia, D.** [US/US]; 1421 S. 248th Street, Des Moines, WA 98198 (US). **DILLON, Davin, C.** [US/US]; 18112 NW Mennonite Drive, Issaquah, WA 98027 (US). **LODES, Michael, L.** [US/US]; 9223 36th Ave. SW, Seattle, WA 98126 (US).

(74) Agent: **PARENT, Annette, S.** et al.; Townsend and Townsend and Crew LLP, Two Embarcadero Center, 8th floor, San Francisco, CA 94111-3834 (US).

(81) Designated States (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

**Published:**
- *With international search report*
- *Before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

*For two letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.*

(54) **Title:** FUSION PROTEINS OF MYCOBACTERIUM TUBERCULOSIS

(57) **Abstract:** The present invention relates to fusion proteins containing at least two *Mycobacterium* species antigens. In particular, it relates to nucleic acids encoding fusion proteins that include two or more individual *M. tuberculosis* antigens, which increase serological sensitivity of sera from individuals infected with tuberculosis, and methods for their use in the diagnosis, treatment, and prevention of tuberculosis infection.

WO 01/24820 A1

# FUSION PROTEINS OF MYCOBACTERIUM TUBERCULOSIS

## CROSS-REFERENCES TO RELATED APPLICATIONS

The present application claims priority to U.S. patent application No. 60/158,338, filed October 7, 1999, and U.S. application No. 60/158,425, filed October 7, 1999, herein each incorporated by reference in its entirety.

This application is also related to U.S. patent application No. 09/056,556, filed April 7, 1998; U.S. patent application No. 09/223,040, filed December 30, 1998; U.S. patent application No. 09/287,849, filed April 7, 1999; and published PCT application No. WO99/51748, filed April 7, 1999 (PCT/US99/07717), herein each incorporated by reference in its entirety.

## STATEMENT AS TO RIGHTS TO INVENTIONS MADE UNDER FEDERALLY SPONSORED RESEARCH AND DEVELOPMENT

Not applicable.

## BACKGROUND OF THE INVENTION

Tuberculosis is a chronic infectious disease caused by infection with *M. tuberculosis* and other Mycobacterium species. It is a major disease in developing countries, as well as an increasing problem in developed areas of the world, with about 8 million new cases and 3 million deaths each year. Although the infection may be asymptomatic for a considerable period of time, the disease is most commonly manifested as an acute inflammation of the lungs, resulting in fever and a nonproductive cough. If untreated, serious complications and death typically result.

Although tuberculosis can generally be controlled using extended antibiotic therapy, such treatment is not sufficient to prevent the spread of the disease. Infected individuals may be asymptomatic, but contagious, for some time. In addition, although compliance with the treatment regimen is critical, patient behavior is difficult to monitor. Some patients do not complete the course of treatment, which can lead to ineffective treatment and the development of drug resistance.

In order to control the spread of tuberculosis, effective vaccination and accurate early diagnosis of the disease are of utmost importance. Currently, vaccination with live bacteria is the most efficient method for inducing protective immunity. The most

common mycobacterium employed for this purpose is Bacillus Calmette-Guerin (BCG), an avirulent strain of *M. bovis*. However, the safety and efficacy of BCG is a source of controversy and some countries, such as the United States, do not vaccinate the general public with this agent.

Diagnosis of tuberculosis is commonly achieved using a skin test, which involves intradermal exposure to tuberculin PPD (protein-purified derivative). Antigen-specific T cell responses result in measurable induration at the injection site by 48-72 hours after injection, which indicates exposure to mycobacterial antigens. Sensitivity and specificity have, however, been a problem with this test, and individuals vaccinated with BCG cannot be distinguished from infected individuals.

While macrophages have been shown to act as the principal effectors of Mycobacterium immunity, T cells are the predominant inducers of such immunity. The essential role of T cells in protection against Mycobacterium infection is illustrated by the frequent occurrence of Mycobacterium infection in AIDS patients, due to the depletion of $CD4^+$ T cells associated with human immunodeficiency virus (HIV) infection. Mycobacterium-reactive $CD4^+$ T cells have been shown to be potent producers of γ-interferon (IFN-γ), which, in turn, has been shown to trigger the anti-mycobacterial effects of macrophages in mice. While the role of IFN-γ in humans is less clear, studies have shown that 1,25-dihydroxy-vitamin D3, either alone or in combination with IFN-γ or tumor necrosis factor-alpha, activates human macrophages to inhibit *M. tuberculosis* infection. Furthermore, it is known that IFN-γ stimulates human macrophages to make 1,25-dihydroxy-vitamin D3. Similarly, interleukin-12 (IL-12) has been shown to play a role in stimulating resistance to *M. tuberculosis* infection. For a review of the immunology of *M. tuberculosis* infection, see Chan & Kaufmann, Tuberculosis: Pathogenesis, Protection and Control (Bloom ed., 1994), and Harrison's Principles of Internal Medicine, volume 1, pp. 1004-1014 and 1019-1023 (14th ed., Fauci et al., eds., 1998).

Accordingly, there is a need for improved diagnostic reagents, and improved methods for diagnosis, preventing and treating tuberculosis.

## SUMMARY OF THE INVENTION

The present invention provides pharmaceutical compositions comprising at least two heterologous antigens, fusion proteins comprising the antigens, and nucleic acids encoding the antigens, where the antigens are from a *Mycobacterium* species from the

tuberculosis complex and other *Mycobacterium* species that cause opportunistic infections in immune compromised patients. The present invention also relates to methods of using the polypeptides and polynucleotides in the diagnosis, treatment and prevention of *Mycobacterium* infection.

The present invention is based, in part, on the inventors' discovery that fusion polynucleotides, fusion polypeptides, or compositions that contain at least two heterologous *M. tuberculosis* coding sequences or antigens are highly antigenic and upon administration to a patient increase the sensitivity of tuberculosis sera. In addition, the compositions, fusion polypeptides and polynucleotides are useful as diagnostic tools in patients that may have been infected with *Mycobacterium*.

In one aspect, the compositions, fusion polypeptides, and nucleic acids of the invention are used in *in vitro* and *in vivo* assays for detecting humoral antibodies or cell-mediated immunity against *M. tuberculosis* for diagnosis of infection or monitoring of disease progression. For example, the polypeptides may be used as an *in vivo* diagnostic agent in the form of an intradermal skin test. The polypeptides may also be used in *in vitro* tests such as an ELISA with patient serum. Alternatively, the nucleic acids, the compositions, and the fusion polypeptides may be used to raise anti-*M. tuberculosis* antibodies in a non-human animal. The antibodies can be used to detect the target antigens *in vivo* and *in vitro*.

In another aspect, the compositions, fusion polypeptides and nucleic acids may be used as immunogens to generate or elicit a protective immune response in a patient. The isolated or purified polynucleotides are used to produce recombinant fusion polypeptide antigens *in vitro*, which are then administered as a vaccine. Alternatively, the polynucleotides may be administered directly into a subject as DNA vaccines to cause antigen expression in the subject, and the subsequent induction of an anti-*M. tuberculosis* immune response. Thus, the isolated or purified *M. tuberculosis* polypeptides and nucleic acids of the invention may be formulated as pharmaceutical compositions for administration to a subject in the prevention and/or treatment of *M. tuberculosis* infection. The immunogenicity of the fusion proteins or antigens may be enhanced by the inclusion of an adjuvant, as well as additional fusion polypeptides, from *Mycobacterium* or other organisms, such as bacterial, viral, mammalian polypeptides. Additional polypeptides may also be included in the compositions, either linked or unlinked to the fusion polypeptide or compositions.

## BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 shows the nucleic acid sequence of a vector encoding TbF14 (SEQ ID NO:89). Nucleotides 5096 to 8594 encode TbF14 (SEQ ID NO:51). Nucleotides 5072 to 5095 encode the eight amino acid His tag (SEQ ID NO:90); nucleotides 5096 to 7315 encode the MTb81 antigen (SEQ ID NO:1); and nucleotides 7316 to 8594 encode the Mo2 antigen (SEQ ID NO:3).

Figure 2 shows the nucleic acid sequence of a vector encoding TbF15 (SEQ ID NO:91). Nucleotides 5096 to 8023 encode the TbF15 fusion protein (SEQ ID NO:53). Nucleotides 5072 to 5095 encode the eight amino acid His tag region (SEQ ID NO:90); nucleotides 5096 to 5293 encode the Ra3 antigen (SEQ ID NO:5); nucleotides 5294 to 6346 encode the 38 kD antigen (SEQ ID NO:7); nucleotides 6347 to 6643 encode the 38-1 antigen (SEQ ID NO:9); and nucleotides 6644 to 8023 encode the FL TbH4 antigen (SEQ ID NO:11).

Figure 3 shows the amino acid sequence of TbF14 (SEQ ID NO:52), including the eight amino acid His tag at the N- terminus.

Figure 4 shows the amino acid sequence of TbF15 (SEQ ID NO:54), including the eight amino acid His tag at the N-terminus.

Figure 5 shows ELISA results using fusion proteins of the invention.

Figure 6 shows the nucleic acid and the predicted amino acid sequences of the entire open reading frame of HTCC#1 FL (SEQ ID NO:13 and 14, respectively).

Figure 7 shows the nucleic acid and predicted amino acid sequences of three fragments of HTCC#1. (a) and (b) show the sequences of two overlapping fragments: an amino terminal half fragment (residues 1 to 223), comprising the first trans-membrane domain (a) and a carboxy terminal half fragment (residues 184 to 392), comprising the last two trans-membrane domains (b); (c) shows a truncated amino-terminal half fragment (residues 1 to 128) devoid of the trans-membrane domain.

Figure 8 shows the nucleic acid and predicted amino acid sequences of a TbRa12-HTCC#1 fusion protein (SEQ ID NO:63 and 64, respectively).

Figure 9a shows the nucleic acid and predicted amino acid sequences of a recombinant HTCC#1 lacking the first trans-membrane domain (deleted of the amino acid residues 150 to 160). Figure 9b shows the nucleic acid and predicted amino acid sequences of 30 overlapping peptides of HTCC#1 used for the T-cell epitope mapping. Figure 9c illustrates the results of the T-cell epitope mapping of HTCC#1. Figure 9d shows the nucleic

acid and predicted amino acid sequences of a deletion construct of HTCC#1 lacking all the trans-membrane domains (deletion of amino acid residues 101 to 203).

Figure 10 shows the nucleic acid and predicted amino acid sequences of the fusion protein HTCC#1(184-392)-TbH9-HTCC#1(1-129) (SEQ ID NO:57 and 58, respectively).

Figure 11 shows the nucleic acid and predicted amino acid sequences of the fusion protein HTCC#1(1-149)-TbH9-HTCC#1(161-392) (SEQ ID NO:59 and 60, respectively).

Figure 12 shows the nucleic acid and predicted amino acid sequences of the fusion protein HTCC#1(184-392)-TbH9-HTCC#1(1-200) (SEQ ID NO:61 and 62, respectively).

Figure 13 shows the nucleotide sequence of *Mycobacterium tuberculosis* antigen MTb59 (SEQ ID NO:49).

Figure 14 shows the amino acid sequence of *Mycobacterium tuberculosis* antigen MTb59 (SEQ ID NO:50).

Figure 15 shows the nucleotide sequence of *Mycobacterium tuberculosis* antigen MTb82 (SEQ ID NO:47).

Figure 16 shows the amino acid sequence of *Mycobacterium tuberculosis* antigen MTb82 (SEQ ID NO:48).

Figure 17 shows the amino acid sequence of *Mycobacterium tuberculosis* the secreted form of antigen DPPD (SEQ ID NO:44).

## DESCRIPTION OF SEQUENCES

SEQ ID NO:1 is the nucleic acid sequence encoding the Mtb81 antigen.

SEQ ID NO:2 is the amino acid sequence of the Mtb81 antigen.

SEQ ID NO:3 is the nucleic acid sequence encoding the Mo2 antigen.

SEQ ID NO:4 is the amino acid sequence of the Mo2 antigen.

SEQ ID NO:5 is the nucleic acid sequence encoding the TbRa3 antigen.

SEQ ID NO:6 is the amino acid sequence of the TbRa3 antigen.

SEQ ID NO:7 is the nucleic acid sequence encoding the 38kD antigen.

SEQ ID NO:8 is the amino acid sequence of the 38kD antigen.

SEQ ID NO:9 is the nucleic acid sequence encoding the Tb38-1 antigen.

SEQ ID NO:10 is the amino acid sequence of the Tb38-1 antigen.

SEQ ID NO:11 is the nucleic acid sequence encoding the full-length (FL) TbH4 antigen.

SEQ ID NO:12 is the amino acid sequence of the FL TbH4 antigen.

SEQ ID NO:13 is the nucleic acid sequence encoding the HTCC#1 (Mtb40) antigen.

SEQ ID NO:14 is the amino acid sequence of the HTCC#1 antigen.

SEQ ID NO:15 is the nucleic acid sequence of an amino terminal half fragment (residues 1 to 223) of HTCC#1, comprising the first trans-membrane domain.

SEQ ID NO:16 is the predicted amino acid sequence of an amino terminal half fragment (residues 1 to 223) of HTCC#1.

SEQ ID NO:17 is the nucleic acid sequence of a carboxy terminal half fragment (residues 184 to 392) of HTCC#1, comprising the last two trans-membrane domains.

SEQ ID NO:18 is the predicted amino acid sequence of a carboxy terminal half fragment (residues 184 to 392) of HTCC#1.

SEQ ID NO:19 is the nucleic acid sequence of a truncated amino-terminal half fragment (residues 1 to 128) of HTCC#1 devoid of the trans-membrane domain.

SEQ ID NO:20 is the predicted amino acid sequence of a truncated amino-terminal half fragment (residues 1 to 128) of HTCC#1.

SEQ ID NO:21 is the nucleic acid sequence of a recombinant HTCC#1 lacking the first trans-membrane domain (deleted of the amino acid residues 150 to 160).

SEQ ID NO:22 is the predicted amino acid sequence of a recombinant HTCC#1 lacking the first trans-membrane domain (deleted of the amino acid residues 150 to 160).

SEQ ID NO:23 is the nucleic acid sequence of a deletion construct of HTCC#1 lacking all the trans-membrane domains (deletion of amino acid residues 101 to 203).

SEQ ID NO:24 is the predicted amino acid sequence of a deletion construct of HTCC#1 lacking all the trans-membrane domains (deletion of amino acid residues 101 to 203).

SEQ ID NO:25 is the nucleic acid sequence encoding the TbH9 (Mtb39A) antigen.

SEQ ID NO:26 is the amino acid sequence of the TbH9 antigen.

SEQ ID NO:27 is the nucleic acid sequence encoding the TbRa12 antigen.

SEQ ID NO:28 is the amino acid sequence of the TbRa12 antigen.

SEQ ID NO:29 is the nucleic acid sequence encoding the TbRa35 (Mtb32A) antigen.

SEQ ID NO:30 is the amino acid sequence of the TbRa35 antigen.

SEQ ID NO:31 is the nucleic acid sequence encoding the MTCC#2 (Mtb41) antigen.

SEQ ID NO:32 is the amino acid sequence of the MTCC#2 antigen.

SEQ ID NO:33 is the nucleic acid sequence encoding the MTI (Mtb9.9A) antigen.

SEQ ID NO:34 is the amino acid sequence of the MTI antigen.

SEQ ID NO:35 is the nucleic acid sequence encoding the MSL (Mtb9.8) antigen.

SEQ ID NO:36 is the amino acid sequence of the MSL antigen.

SEQ ID NO:37 is the nucleic acid sequence encoding the DPV (Mtb8.4) antigen.

SEQ ID NO:38 is the amino acid sequence of the DPV antigen.

SEQ ID NO:39 is the nucleic acid sequence encoding the DPEP antigen.

SEQ ID NO:40 is the amino acid sequence of the DPEP antigen.

SEQ ID NO:41 is the nucleic acid aequence encoding the Erd14 (Mtb16) antigen.

SEQ ID NO:42 is the amino acid sequence of the Erd14 antigen.

SEQ ID NO:43 is the nucleic acid sequence encoding the DPPD antigen.

SEQ ID NO:44 is the amino acid sequence of the DPPD antigen.

SEQ ID NO:45 is the nucleic acid sequence encoding the ESAT-6 antigen.

SEQ ID NO:46 is the amino acid sequence of the ESAT-6 antigen.

SEQ ID NO:47 is the nucleic acid sequence encoding the Mtb82 (Mtb867) antigen.

SEQ ID NO:48 is the amino acid sequence of the Mtb82 antigen.

SEQ ID NO:49 is the nucleic acid sequence encoding the Mtb59 (Mtb403) antigen.

SEQ ID NO:50 is the amino acid sequence of the Mtb59 antigen.

SEQ ID NO:51 is the nucleic acid sequence encoding the TbF14 fusion protein.

SEQ ID NO:52 is the amino acid sequence of the TbF14 fusion protein.

SEQ ID NO:53 is the nucleic acid sequence encoding the TbF15 fusion protein.

SEQ ID NO:54 is the amino acid sequence of the TbF15 fusion protein.

SEQ ID NO:55 is the nucleic acid sequence of the fusion protein HTCC#1(FL)-TbH9(FL).

SEQ ID NO:56 is the amino acid sequence of the fusion protein HTCC#1(FL)-TbH9(FL).

SEQ ID NO:57 is the nucleic acid sequence of the fusion protein HTCC#1(184-392)-TbH9-HTCC#1(1-129).

SEQ ID NO:58 is the predicted amino acid of the fusion protein HTCC#1(184-392)-TbH9-HTCC#1(1-129).

SEQ ID NO:59 is the nucleic acid sequence of the fusion protein HTCC#1(1-149)-TbH9-HTCC#1(161-392).

SEQ ID NO:60 is the predicted amino acid sequence of the fusion protein HTCC#1(1-149)-TbH9-HTCC#1(161-392).

SEQ ID NO:61 is the nucleic acid sequence of the fusion protein HTCC#1(184-392)-TbH9-HTCC#1(1-200).

SEQ ID NO:62 is the predicted amino acid sequence of the fusion protein HTCC#1(184-392)-TbH9-HTCC#1(1-200).

SEQ ID NO:63 is the nucleic acid sequence of the TbRa12-HTCC#1 fusion protein.

SEQ ID NO:64 is the predicted amino acid sequence of the TbRa12-HTCC#1 fusion protein.

SEQ ID NO:65 is the nucleic acid sequence of the TbF (TbRa3, 38kD, Tb38-1) fusion protein.

SEQ ID NO:66 is the predicted amino acid sequence of the TbF fusion protein.

SEQ ID NO:67 is the nucleic acid sequence of the TbF2 (TbRa3, 38kD, Tb38-1, DPEP) fusion protein.

SEQ ID NO:68 is the predicted amino acid sequence of the TbF2 fusion protein.

SEQ ID NO:69 is the nucleic acid sequence of the TbF6 (TbRa3, 38kD, Tb38-1, TbH4) fusion protein.

SEQ ID NO:70 is the predicted amino acid sequence of the TbF6 fusion protein.

SEQ ID NO:71 is the nucleic acid sequence of the TbF8 (38kD-linker-DPEP) fusion protein.

SEQ ID NO:72 is the predicted amino acid sequence of the TbF8 fusion protein.

SEQ ID NO:73 is the nucleic acid sequence of the Mtb36F (Erd14-DPV-MTI) fusion protein.

SEQ ID NO:74 is the predicted amino acid sequence of the Mtb36F fusion protein.

SEQ ID NO:75 is the nucleic acid sequence of the Mtb88F (Erd14-DPV-MTI-MSL-MTCC#2) fusion protein.

SEQ ID NO:76 is the predicted amino acid sequence of the Mtb88F fusion protein.

SEQ ID NO:77 is the nucleic acid sequence of the Mtb46F (Erd14-DPV-MTI-MSL) fusion protein.

SEQ ID NO:78 is the predicted amino acid sequence of the Mtb46F fusion protein.

SEQ ID NO:79 is the nucleic acid sequence of the Mtb71F (DPV-MTI-MSL-MTCC#2) fusion protein.

SEQ ID NO:80 is the predicted amino acid sequence of the Mtb71F fusion protein.

SEQ ID NO:81 is the nucleic acid sequence of the Mtb31F (DPV-MTI-MSL) fusion protein.

SEQ ID NO:82 is the predicted amino acid sequence of the Mtb31F fusion protein.

SEQ ID NO:83 is the nucleic acid sequence of the Mtb61F (TbH9-DPV-MTI) fusion protein.

SEQ ID NO:84 is the predicted amino acid sequence of the Mtb61F fusion protein.

SEQ ID NO:85 is the nucleic acid sequence of the Ra12-DPPD (Mtb24F) fusion protein.

SEQ ID NO:86 is the predicted amino acid sequence of the Ra12-DPPD fusion protein.

SEQ ID NO:87 is the nucleic acid sequence of the Mtb72F (TbRa12-TbH9-TbRa35) fusion protein.

SEQ ID NO:88 is the predicted amino acid sequence of the Mtb72F fusion protein.

SEQ ID NO:89 is the nucleic acid sequence of the Mtb59F (TbH9-TbRa35) fusion protein.

SEQ ID NO:90 is the predicted amino acid sequence of the Mtb59F fusion protein.

SEQ ID NO:91 is the nucleic acid sequence of a vector encoding TbF14.

SEQ ID NO:92 is the nucleotide sequence of the region spanning nucleotides 5072 to 5095 of SEQ ID NO:91 encoding the eight amino acid His tag.

SEQ ID NO:93 is the nucleic acid sequence of a vector encoding TbF15.

SEQ ID NO:94-123 are the nucleic acid sequences of 30 overlapping peptides of HTCC#1 used for the T-cell epitope mapping.

SEQ ID NO:124-153 are the predicted amino acid sequences of 30 overlapping peptides of HTCC#1 used for the T-cell epitope mapping.

## DETAILED DESCRIPTION OF THE INVENTION

### I. INTRODUCTION

The present invention relates to compositions comprising antigen compositions and fusion polypeptides useful for the diagnosis and treatment of *Mycobacterium* infection, polynucleotides encoding such antigens, and methods for their use. The antigens of the present invention are polypeptides or fusion polypeptides of *Mycobacterium* antigens and immunogenic fragments thereof. More specifically, the compositions of the present invention comprise at least two heterologous polypeptides of a *Mycobacterium* species of the tuberculosis complex, *e.g.*, a species such as *M. tuberculosis*, *M. bovis*, or *M. africanum*, or a *Mycobacterium* species that is environmental or opportunistic and that causes opportunistic infections such as lung infections in immune compromised hosts (*e.g.*, patients with AIDS), *e.g.*, *BCG*, *M. avium*, *M. intracellulare*, *M. celatum*, *M. genavense*, *M. haemophilum*, *M. kansasii*, *M. simiae*, *M. vaccae*, *M. fortuitum*, and *M. scrofulaceum* (*see, e.g., Harrison's Principles of Internal Medicine*, volume 1, pp. 1004-1014 and 1019-1023 (14th ed., Fauci *et al.*, eds., 1998). The inventors of the present application surprisingly discovered that compositions and fusion proteins comprising at least two

heterologous *Mycobacterium* antigens, or immunogenic fragments thereof, where highly antigenic. These compositions, fusion polypeptides, and the nucleic acids that encode them are therefore useful for eliciting protective response in patients, and for diagnostic applications.

The antigens of the present invention may further comprise other components designed to enhance the antigenicity of the antigens or to improve these antigens in other aspects, for example, the isolation of these antigens through addition of a stretch of histidine residues at one end of the antigen. The compositions, fusion polypeptides, and nucleic acids of the invention can comprise additional copies of antigens, or additional heterologous polypeptides from *Mycobacterium* species, such as, *e.g.*, MTb81, Mo2, TbRa3, 38 kD (with the N-terminal cysteine residue), Tb38-1, FL TbH4, HTCC#1, TbH9, MTCC#2, MTI, MSL, TbRa35, DPV, DPEP, Erd14, TbRa12, DPPD, MTb82, MTb59, ESAT-6, MTB85 complex, or α-crystalline. Such fusion polypeptides are also referred to as polyproteins. The compositions, fusion polypeptides, and nucleic acids of the invention can also comprise additional polypeptides from other sources. For example, the compositions and fusion proteins of the invention can include polypeptides or nucleic acids encoding polypeptides, wherein the polypeptide enhances expression of the antigen, *e.g.*, NS1, an influenza virus protein, or an immunogenic portion thereof (*see, e.g.*, WO99/40188 and WO93/04175). The nucleic acids of the invention can be engineered based on codon preference in a species of choice, *e.g.*, humans.

The compositions of the invention can be naked DNA, or the compositions, *e.g.*, polypeptides, can also comprise adjuvants such as, for example, AS2, AS2', AS2'', AS4, AS6, ENHANZYN (Detox), MPL, QS21, CWS, TDM, AGPs, CPG, Leif, saponin, and saponin mimetics, and derivatives thereof.

In one aspect, the compositions and fusion proteins of the invention are composed of at least two antigens selected from the group consisting of an MTb81 antigen or an immunogenic fragment thereof from a *Mycobacterium* species of the tuberculosis complex, and an Mo2 antigen or an immunogenic fragment thereof from a *Mycobacterium* species of the tuberculosis complex. In one embodiment, the compositions of the invention comprise the TbF14 fusion protein. The complete nucleotide sequence encoding TbF14 is set forth in SEQ ID NO:51, and the amino acid sequence of TbF14 is set forth in SEQ ID NO:52.

In another aspect, the compositions and fusion proteins of the invention are composed of at least four antigens selected from the group consisting of a TbRa3 antigen or an immunogenic fragment thereof from a *Mycobacterium* species of the tuberculosis

complex, a 38 kD antigen or an immunogenic fragment thereof from a *Mycobacterium* species of the tuberculosis complex, a Tb38-1 antigen or an immunogenic fragment thereof from a *Mycobacterium* species of the tuberculosis complex, and a FL TbH4 antigen or an immunogenic fragment thereof from a *Mycobacterium* species of the tuberculosis complex. In one embodiment, the compositions of the invention comprise the TbF15 fusion protein. The nucleic acid and amino acid sequences of TbF15 are set forth in SEQ ID NO:53 and 54, respectively.

In another aspect, the compositions and fusion proteins of the invention are composed of at least two antigens selected from the group consisting of an HTCC#1 antigen or an immunogenic fragment thereof from a *Mycobacterium* species of the tuberculosis complex, and a TbH9 antigen or an immunogenic fragment thereof from a *Mycobacterium* species of the tuberculosis complex. In one embodiment, the compositions of the invention comprise the HTCC#1(FL)-TbH9(FL) fusion protein. The nucleic acid and amino acid sequences of HTCC#1-TbH9 are set forth in SEQ ID NO:55 and 56, respectively. In another embodiment, the compositions of the invention comprise the fusion protein HTCC#1(184-392)/TbH9/HTCC#1(1-129). The nucleic acid and amino acid sequences of HTCC#1(184-392)/TbH9/HTCC#1(1-129) are set forth in SEQ ID NO:57 and 58, respectively. In yet another embodiment, the compositions of the invention comprise the fusion protein HTCC#1(1-149)/TbH9/HTCC#1(161-392), having the nucleic acid and amino acid sequences set forth in SEQ ID NO:59 and 60, respectively. In still another embodiment, the compositions of the invention comprise the fusion protein HTCC#1(184-392)/TbH9/HTCC#1(1-200), having the nucleic acid and amino acid sequences set forth in SEQ ID NO:61 and 62, respectively.

In a different aspect, the compositions and fusion proteins of the invention are composed of at least two antigens selected from the group consisting of an HTCC#1 antigen or an immunogenic fragment thereof from a *Mycobacterium* species of the tuberculosis complex, and a TbRa12 antigen or an immunogenic fragment thereof from a *Mycobacterium* species of the tuberculosis complex. In one embodiment, the compositions of the invention comprise the fusion protein TbRa12-HTCC#1. The nucleic acid and amino acid sequences of the TbRa12-HTCC#1 fusion protein are set forth in SEQ ID NO:63 and 64, respectively.

In yet another aspect, the compositions and fusion proteins of the invention are composed of at least two antigens selected from the group consisting of a TbH9 (MTB39) antigen or an immunogenic fragment thereof from a *Mycobacterium* species of the tuberculosis complex, and a TbRa35 (MTB32A) antigen or an immunogenic fragment thereof

from a *Mycobacterium* species of the tuberculosis complex. In one embodiment, the antigens are selected from the group consisting of a TbH9 (MTB39) antigen or an immunogenic fragment thereof from a *Mycobacterium* species of the tuberculosis complex, and a polypeptide comprising at least 205 amino acids of the N-terminus of a TbRa35 (MTB32A) antigen from a *Mycobacterium* species of the tuberculosis complex. In another embodiment, the antigens are selected from the group consisting of a TbH9 (MTB39) antigen or an immunogenic fragment thereof from a *Mycobacterium* species of the tuberculosis complex, a polypeptide comprising at least 205 amino acids of the N-terminus of a TbRa35 (MTB32A) antigen from a *Mycobacterium* species of the tuberculosis complex, and a polypeptide comprising at least about 132 amino acids from the C-terminus of a TbRa35 (MTB32A) antigen from a *Mycobacterium* species of the tuberculosis complex.

In yet another embodiment, the compositions of the invention comprise the Mtb59F fusion protein. The nucleic acid and amino acid sequences of the Mtb59F fusion protein are set forth in SEQ ID NO:89 and 90, respectively, as well as in the U.S. patent application No. 09/287,849 and in the PCT/US99/07717 application. In another embodiment, the compositions of the invention comprise the Mtb72F fusion protein having the nucleic acid and amino acid sequences set forth in SEQ ID NO:87 and 88, respectively. The Mtb72F fusion protein is also disclosed in the U.S. patent application Nos. 09/223,040 and 09/223,040; and in the PCT/US99/07717 application.

In yet another aspect, the compositions and fusion proteins of the invention comprise at least two antigens selected from the group consisting of MTb81, Mo2, TbRa3, 38kD, Tb38-1 (MTb11), FL TbH4, HTCC#1 (Mtb40), TbH9, MTCC#2 (Mtb41), DPEP, DPPD, TbRa35, TbRa12, MTb59, MTb82, Erd14 (Mtb16), FL TbRa35 (Mtb32A), DPV (Mtb8.4), MSL (Mtb9.8), MTI (Mtb9.9A, also known as MTI-A), ESAT-6, α-crystalline, and 85 complex, or an immunogenic fragment thereof from a *Mycobacterium* species of the tuberculosis complex.

In another aspect, the fusion proteins of the invention are:

TbRa3-38 kD-Tb38-1 (TbF), the sequence of which is disclosed in SEQ ID NO:65 (DNA) and SEQ ID NO:66 (protein), as well as in the U.S. patent application Nos. 08/818,112; 08/818,111; and 09/056,556; and in the WO98/16646 and WO98/16645 applications;

TbRa3-38kD-Tb38-1-DPEP (TbF2), the sequence of which is disclosed in SEQ ID NO:67 (DNA) and SEQ ID NO:68 (protein), and in the U.S. patent application Nos. 08/942,578; 08/942,341; 09/056,556; and in the WO98/16646 and WO98/16645 applications;

TbRa3-38kD-Tb38-1-TBH4 (TbF6), the sequence of which is disclosed in SEQ ID NO:69 (DNA) and SEQ ID NO:70 (protein) in the U.S. patent application Nos. 08/072,967; 09/072,596; and in the PCT/US99/03268 and PCT/US99/03265 applications;

38kD-Linker-DPEP (TbF8), the sequence of which is disclosed in SEQ ID NO:71 (DNA) and SEQ ID NO:72 (protein), and in the U.S. patent application Nos. 09/072,967 and 09/072,596; as well as in the PCT/US99/03268 and PCT/US99/03265 applications;

Erd14-DPV-MTI (MTb36F), the sequence of which is disclosed in SEQ ID NO:73 (DNA), SEQ ID NO:74 (protein), as well as in the U.S. patent application Nos. 09/223,040 and No. 09/287,849; and in the PCT/US99/07717 application;

Erd14-DPV-MTI-MSL-MTCC#2 (MTb88F), the sequence of which is disclosed in SEQ ID NO:75 (cDNA) and SEQ ID NO:76 (protein), as well as in the U.S. patent application No. 09/287,849 and in the PCT/US99/07717 application;

Erd14-DPV-MTI-MSL (MTb46F), the sequence of which is disclosed in SEQ ID NO:77 (cDNA) and SEQ ID NO:78 (protein), and in the U.S. patent application No. 09/287,849 and in the PCT/US99/07717 application;

DPV-MTI-MSL-MTCC#2 (MTb71F), the sequence of which is disclosed in SEQ ID NO:79 (cDNA) and SEQ ID NO:80 (protein), as well as in the U.S. patent application No. 09/287,849 and in the PCT/US99/07717 application;

DPV-MTI-MSL (MTb31F), the sequence of which is disclosed in SEQ ID NO:81 (cDNA) and SEQ ID NO:82 (protein), and in the U.S. patent application No. 09/287,849 and in the PCT/US99/07717 application;

TbH9-DPV-MTI (MTb61F), the sequence of which is disclosed in SEQ ID NO:83 (cDNA) and SEQ ID NO:84 (protein) (*see*, also, U.S. patent application No. 09/287,849 and PCT/US99/07717 application);

Ra12-DPPD (MTb24F), the sequence of which is disclosed in SEQ ID NO:85 (cDNA) and SEQ ID NO:86 (protein), as well as in the U.S. patent application No. 09/287,849 and in the PCT/US99/07717 application.

In the nomenclature of the application, TbRa35 refers to the N-terminus of MTB32A (TbRa35FL), comprising at least about the first 205 amino acids of MTB32A from *M. tuberculosis*, or the corresponding region from another *Mycobacterium* species. TbRa12

refers to the C-terminus of MTB32A (TbRa35FL), comprising at least about the last 132 amino acids from MTB32A from *M. tuberculosis*, or the corresponding region from another *Mycobacterium* species.

The following provides sequences of some individual antigens used in the compositions and fusion proteins of the invention:

Mtb81, the sequence of which is disclosed in SEQ ID NO:1 (DNA) and SEQ ID NO:2 (predicted amino acid).

Mo2, the sequence of which is disclosed in SEQ ID NO:3 (DNA) and SEQ ID NO:4 (predicted amino acid).

Tb38-1 or 38-1 (MTb11), the sequence of which is disclosed in SEQ ID NO:9 (DNA) and SEQ ID NO:10 (predicted amino acid), and is also disclosed in the U.S. patent application Nos. 09/072,96; 08/523,436; 08/523,435; 08/818,112; and 08/818,111; and in the WO97/09428 and WO97/09429 applications;

TbRa3, the sequence of which is disclosed in SEQ ID NO:5 (DNA) and SEQ ID NO:6 (predicted amino acid sequence) (*see*, also, WO 97/09428 and WO97/09429 applications);

38 kD, the sequence of which is disclosed in SEQ ID NO:7 (DNA) and SEQ ID NO:8 (predicted amino acid sequence), as well as in the U.S. patent application No. 09/072,967. 38 kD has two alternative forms, with and without the N-terminal cysteine residue;

DPEP, the sequence of which is disclosed in SEQ ID NO:39 (DNA) and SEQ ID NO:40 (predicted amino acid sequence), and in the WO97/09428 and WO97/09429 publications;

TbH4, the sequence of which is disclosed as SEQ ID NO:11 (DNA) and SEQ ID NO:12 (predicted amino acid sequence) (*see*, also, WO97/09428 and WO97/09429 publications);

Erd14 (MTb16), the cDNA and amino acids sequences of which are disclosed in SEQ ID NO:41 (DNA) and 42 (predicted amino acid), and in Verbon *et al., J. Bacteriology* 174:1352-1359 (1992);

DPPD, the sequence of which is disclosed in SEQ ID NO:43 (DNA) and SEQ ID NO:44 (predicted amino acid sequence), and in the PCT/US99/03268 and PCT/US99/03265 applications. The secreted form of DPPD is shown herein in Figure 12;

MTb82 (MTb867), the sequence of which is disclosed in SEQ ID NO:47 (DNA) and SEQ ID NO:48 (predicted amino acid sequence), and in Figures 8 (DNA) and 9 (amino acid);

MTb59 (MTb403) , the sequence of which is disclosed in SEQ ID NO:49 (DNA) and SEQ ID NO:50 (predicted amino acid sequence), and in Figures 10 (DNA) and 11 (amino acid);

TbRa35 FL (MTB32A), the sequence of which is disclosed as SEQ ID NO:29 (cDNA) and SEQ ID NO:30 (protein), and in the U.S. patent application Nos. 08/523,436, 08/523,435; 08/658,800; 08/659,683; 08/818,112; 09/056,556; and 08/818,111; as well as in the WO97/09428 and WO97/09429 applications; *see also* Skeiky *et al., Infection and Immunity* 67:3998-4007 (1999);

TbRa12, the C-terminus of MTB32A (TbRa35FL), comprising at least about the last 132 amino acids from MTB32A from *M. tuberculosis*, the sequence of which is disclosed as SEQ ID NO:27 (DNA) and SEQ ID NO:28 (predicted amino acid sequence) (*see*, also, U.S. patent application No. 09/072,967; and WO97/09428 and WO97/09429 publications);

TbRa35, the N-terminus of MTB32A (TbRa35FL), comprising at least about the first 205 amino acids of MTB32A from *M. tuberculosis*, the nucleotide and amino acid sequence of which is disclosed in Figure 4;

TbH9 (MTB39), the sequence of which is disclosed in SEQ ID NO:25 (cDNA full length) and SEQ ID NO:26 (protein full length), as well as in the U.S. patent application Nos. 08/658,800; 08/659,683; 08/818,112; 08/818,111; and 09/056,559; and in the WO97/09428 and WO97/09429 applications.

HTCC#1 (MTB40), the sequence of which is disclosed in SEQ ID NO:13 (DNA) and SEQ ID NO:14 (amino acid), as well as in the U.S. patent application Nos. 09/073,010; and 09/073,009; and in the PCT/US98/10407 and PCT/US98/10514 applications;

MTCC#2 (MTB41), the sequence of which is disclosed in SEQ ID NO:31 (DNA) and SEQ ID NO:32 (amino acid), as well as in the U.S. patent application Nos. 09/073,010; and 09/073,009; and in the WO98/53075 and WO98/53076 publications;

MTI (Mtb9.9A), the sequence of which is disclosed in SEQ ID NO:33 (DNA) and SEQ ID NO:34 (amino acid), as well as in the U.S. patent application Nos. 09/073,010; and 09/073,009; and in the WO98/53075 and WO98/53076 publications;

MSL (Mtb9.8), the sequence of which is disclosed in SEQ ID NO:35 (DNA) and SEQ ID NO:36 (amino acid), as well as in the U.S. patent application Nos. 09/073,010; and 09/073,009; and in the WO98/53075 and WO98/53076 publications;

DPV (Mtb8.4), the sequence of which is disclosed in SEQ ID NO:37 (DNA) and SEQ ID NO:38 (amino acid), and in the U.S. patent application Nos. 08/658,800; 08/659,683; 08/818,111; 08/818,112; as well as in the WO97/09428 and WO97/09429 publications;

ESAT-6 (Mtb8.4), the sequence of which is disclosed in SEQ ID NO:45 (DNA) and SEQ ID NO:46 (amino acid), and in the U.S. patent application Nos. 08/658,800; 08/659,683; 08/818,111; 08/818,112; as well as in the WO97/09428 and WO97/09429 publications;

The following provides sequences of some additional antigens used in the compositions and fusion proteins of the invention:

α-crystalline antigen, the sequence of which is disclosed in Verbon *et al., J. Bact.* 174:1352-1359 (1992);

85 complex antigen, the sequence of which is disclosed in Content *et al., Infect. & Immunol.* 59:3205-3212 (1991).

Each of the above sequences is also disclosed in Cole *et al. Nature* 393:537 (1998) and can be found at, e.g., http://www.sanger.ac.uk and http:/www.pasteur.fr/mycdb/.

The above sequences are disclosed in U.S. patent applications Nos. 08/523,435; 08/523,436; 08/658,800; 08/659,683; 08/818,111; 08/818,112; 08/942,341; 08/942,578; 08/858,998; 08/859,381; 09/056,556; 09/072,596; 09/072,967; 09/073,009; 09/073,010; 09/223,040; 09/287,849; and in PCT patent applications PCT/US99/03265, PCT/US99/03268; PCT/US99/07717; WO97/09428; WO97/09429; WO98/16645; WO98/16646; WO98/53075; and WO98/53076, each of which is herein incorporated by reference.

The antigens described herein include polymorphic variants and conservatively modified variations, as well as inter-strain and interspecies *Mycobacterium* homologs. In addition, the antigens described herein include subsequences or truncated sequences. The fusion proteins may also contain additional polypeptides, optionally heterologous peptides from *Mycobacterium* or other sources. These antigens may be modified, for example, by adding linker peptide sequences as described below. These linker peptides may be inserted between one or more polypeptides which make up each of the fusion proteins.

## II. DEFINITIONS

"Fusion polypeptide" or "fusion protein" refers to a protein having at least two heterologous *Mycobacterium* sp. polypeptides covalently linked, either directly or via an amino acid linker. The polypeptides forming the fusion protein are typically linked C-terminus to N-terminus, although they can also be linked C-terminus to C-terminus, N-terminus to N-terminus, or N-terminus to C-terminus. The polypeptides of the fusion protein can be in any order. This term also refers to conservatively modified variants, polymorphic variants, alleles, mutants, subsequences, and interspecies homologs of the antigens that make up the fusion protein. *Mycobacterium tuberculosis* antigens are described in Cole *et al.*, *Nature* 393:537 (1998), which discloses the entire *Mycobacterium tuberculosis* genome. The complete sequence of *Mycobacterium tuberculosis* can also be found at http://www.sanger.ac.uk and at http://www.pasteur.fr/mycdb/ (MycDB). Antigens from other *Mycobacterium* species that correspond to *M. tuberculosis* antigens can be identified, *e.g.*, using sequence comparison algorithms, as described herein, or other methods known to those of skill in the art, *e.g.*, hybridization assays and antibody binding assays.

The term "TbF14" refers to a fusion protein having at least two antigenic, heterologous polypeptides from *Mycobacterium* fused together. The two peptides are referred to as MTb81 and Mo2. This term also refers to a fusion protein having polymorphic variants, alleles, mutants, fragments, and interspecies homologs of MTb81 and Mo2. A nucleic acid encoding TbF14 specifically hybridizes under highly stringent hybridization conditions to SEQ ID NO:1 and 3, which individually encode the MTb81 and Mo2 antigens, respectively, and alleles, polymorphic variants, interspecies homologs, subsequences, and conservatively modified variants thereof. A TbF14 fusion polypeptide specifically binds to antibodies raised against MTb81 and Mo2, and alleles, polymorphic variants, interspecies homologs, subsequences, and conservatively modified variants thereof (optionally including an amino acid linker). The antibodies are polyclonal or monoclonal. Optionally, the TbF14 fusion polypeptide specifically binds to antibodies raised against the fusion junction of MTb81 and Mo2, which antibodies do not bind to MTb81 or Mo2 individually, *i.e.*, when they are not part of a fusion protein. The individual polypeptides of the fusion protein can be in any order. In some embodiments, the individual polypeptides are in order (N- to C- terminus) from large to small. Large antigens are approximately 30 to 150 kD in size, medium antigens are approximately 10 to 30 kD in

size, and small antigens are approximately less than 10 kD in size. The sequence encoding the individual polypeptide may be, *e.g.*, a fragment such as an individual CTL epitope encoding about 8 to 9 amino acids. The fragment may also include multiple epitopes. The fragment may also represent a larger part of the antigen sequence, *e.g.*, about 50% or more of MTb81 and Mo2.

TbF14 optionally comprises additional polypeptides, optionally heterologous polypeptides, fused to MTb81 and Mo2, optionally derived from *Mycobacterium* as well as other sources, such as viral, bacterial, eukaryotic, invertebrate, vertebrate, and mammalian sources. As described herein, the fusion protein can also be linked to other molecules, including additional polypeptides.

The term "TbF15" refers to a fusion protein having at least four antigenic, heterologous polypeptides from *Mycobacterium* fused together. The four peptides are referred to as TbRa3, 38 kD, Tb38-1 (with the N-terminal cysteine), and FL TbH4. This term also refers to a fusion protein having polymorphic variants, alleles, mutants, and interspecies homologs of TbRa3, 38 kD, Tb38-1, and FL TbH4. A nucleic acid encoding TbF15 specifically hybridizes under highly stringent hybridization conditions to SEQ ID NO:5, 7, 9 and 11, individually encoding TbRa3, 38 kD, Tb38-1 and FL TbH4, respectively, and alleles, fragments, polymorphic variants, interspecies homologs, subsequences, and conservatively modified variants thereof. A TbF15 fusion polypeptide specifically binds to antibodies raised against TbRa3, 38 kD, Tb38-1, and FL TbH4 and alleles, polymorphic variants, interspecies homologs, subsequences, and conservatively modified variants thereof (optionally including an amino acid linker). The antibodies are polyclonal or monoclonal. Optionally, the TbF15 fusion polypeptide specifically binds to antibodies raised against the fusion junction of TbRa3, 38 kD, Tb38-1, and FL TbH4, which antibodies do not bind to TbRa3, 38 kD, Tb38-1, and FL TbH4 individually, *i.e.*, when they are not part of a fusion protein. The polypeptides of the fusion protein can be in any order. In some embodiments, the individual polypeptides are in order (N- to C- terminus) from large to small. Large antigens are approximately 30 to 150 kD in size, medium antigens are approximately 10 to 30 kD in size, and small antigens are approximately less than 10 kD in size. The sequence encoding the individual polypeptide may be as small as, *e.g.*, a fragment such as an individual CTL epitope encoding about 8 to 9 amino acids. The fragment may also include multiple epitopes. The fragment may also represent a larger

part of the antigen sequence, *e.g.*, about 50% or more of TbRa3, 38 kD, Tb38-1, and FL TbH4.

TbF15 optionally comprises additional polypeptides, optionally heterologous polypeptides, fused to TbRa3, 38 kD, Tb38-1, and FL TbH4, optionally derived from *Mycobacterium* as well as other sources such as viral, bacterial, eukaryotic, invertebrate, vertebrate, and mammalian sources. As described herein, the fusion protein can also be linked to other molecules, including additional polypeptides. The compositions of the invention can also comprise additional polypeptides that are unlinked to the fusion proteins of the invention. These additional polypeptides may be heterologous or homologous polypeptides.

The "HTCC#1(FL)-TbH9(FL)," "HTCC#1(184-392)/TbH9/HTCC#1(1-129)," "HTCC#1(1-149)/TbH9/HTCC#1(161-392)," and "HTCC#1(184-392)/TbH9/HTCC#1(1-200)" fusion proteins refer to fusion proteins comprising at least two antigenic, heterologous polypeptides from *Mycobacterium* fused together. The two peptides are referred to as HTCC#1 and TbH9. This term also refers to fusion proteins having polymorphic variants, alleles, mutants, and interspecies homologs of HTCC#1 and TbH9. A nucleic acid encoding HTCC#1-TbH9, HTCC#1(184-392)/TbH9/HTCC#1(1-129), HTCC#1(1-149)/TbH9/HTCC#1(161-392), or HTCC#1(184-392)/TbH9/HTCC#1(1-200) specifically hybridizes under highly stringent hybridization conditions to SEQ ID NO:13 and 25, individually encoding HTCC#1 and TbH9, respectively, and alleles, fragments, polymorphic variants, interspecies homologs, subsequences, and conservatively modified variants thereof. A HTCC#1(FL)-TbH9(FL), HTCC#1(184-392)/TbH9/HTCC#1(1-129), HTCC#1(1-149)/TbH9/HTCC#1(161-392), or HTCC#1(184-392)/TbH9/HTCC#1(1-200) fusion polypeptide specifically binds to antibodies raised against HTCC#1 and TbH9, and alleles, polymorphic variants, interspecies homologs, subsequences, and conservatively modified variants thereof (optionally including an amino acid linker). The antibodies are polyclonal or monoclonal. Optionally, the HTCC#1(FL)-TbH9(FL), HTCC#1(184-392)/TbH9/HTCC#1(1-129), HTCC#1(1-149)/TbH9/HTCC#1(161-392), or HTCC#1(184-392)/TbH9/HTCC#1(1-200) fusion polypeptide specifically binds to antibodies raised against the fusion junction of the antigens, which antibodies do not bind to the antigens individually, *i.e.*, when they are not part of a fusion protein. The polypeptides of the fusion protein can be in any order. In some embodiments, the individual polypeptides are in order

(N- to C- terminus) from large to small. Large antigens are approximately 30 to 150 kD in size, medium antigens are approximately 10 to 30 kD in size, and small antigens are approximately less than 10 kD in size. The sequence encoding the individual polypeptide may be as small as, *e.g.*, a fragment such as an individual CTL epitope encoding about 8 to 9 amino acids. The fragment may also include multiple epitopes. The fragment may also represent a larger part of the antigen sequence, *e.g.*, about 50% or more (*e.g.*, full-length) of HTCC#1 and TbH9.

HTCC#1(FL)-TbH9(FL), HTCC#1(184-392)/TbH9/HTCC#1(1-129), HTCC#1(1-149)/TbH9/HTCC#1(161-392), and HTCC#1(184-392)/TbH9/HTCC#1(1-200) optionally comprise additional polypeptides, optionally heterologous polypeptides, fused to HTCC#1 and TbH9, optionally derived from *Mycobacterium* as well as other sources such as viral, bacterial, eukaryotic, invertebrate, vertebrate, and mammalian sources. As described herein, the fusion protein can also be linked to other molecules, including additional polypeptides. The compositions of the invention can also comprise additional polypeptides that are unlinked to the fusion proteins of the invention. These additional polypeptides may be heterologous or homologous polypeptides.

The term "TbRa12-HTCC#1" refers to a fusion protein having at least two antigenic, heterologous polypeptides from *Mycobacterium* fused together. The two peptides are referred to as TbRa12 and HTCC#1. This term also refers to a fusion protein having polymorphic variants, alleles, mutants, and interspecies homologs of TbRa12 and HTCC#1. A nucleic acid encoding "TbRa12-HTCC#1" specifically hybridizes under highly stringent hybridization conditions to SEQ ID NO:27 and 13, individually encoding TbRa12 and HTCC#1, respectively, and alleles, fragments, polymorphic variants, interspecies homologs, subsequences, and conservatively modified variants thereof. A "TbRa12-HTCC#1" fusion polypeptide specifically binds to antibodies raised against TbRa12 and HTCC#1 and alleles, polymorphic variants, interspecies homologs, subsequences, and conservatively modified variants thereof (optionally including an amino acid linker). The antibodies are polyclonal or monoclonal. Optionally, the "TbRa12-HTCC#1" fusion polypeptide specifically binds to antibodies raised against the fusion junction of TbRa12 and HTCC#1, which antibodies do not bind to TbRa12 and HTCC#1 individually, *i.e.*, when they are not part of a fusion protein. The polypeptides of the fusion protein can be in any order. In some embodiments, the individual polypeptides are in order (N- to C- terminus)

from large to small. Large antigens are approximately 30 to 150 kD in size, medium antigens are approximately 10 to 30 kD in size, and small antigens are approximately less than 10 kD in size. The sequence encoding the individual polypeptide may be as small as, *e.g.*, a fragment such as an individual CTL epitope encoding about 8 to 9 amino acids. The fragment may also include multiple epitopes. The fragment may also represent a larger part of the antigen sequence, *e.g.*, about 50% or more of TbRa12 and HTCC#1.

"TbRa12-HTCC#1" optionally comprises additional polypeptides, optionally heterologous polypeptides, fused to TbRa12 and HTCC#1, optionally derived from *Mycobacterium* as well as other sources such as viral, bacterial, eukaryotic, invertebrate, vertebrate, and mammalian sources. As described herein, the fusion protein can also be linked to other molecules, including additional polypeptides. The compositions of the invention can also comprise additional polypeptides that are unlinked to the fusion proteins of the invention. These additional polypeptides may be heterologous or homologous polypeptides.

The term "Mtb72F" and "Mtb59F" refer to fusion proteins of the invention which hybridize under stringent conditions to at least two nucleotide sequences set forth in SEQ ID NO:25 and 29, individually encoding the Tbh9 (MTB39) and Ra35 (MTB32A) antigens. The polynucleotide sequences encoding the individual antigens of the fusion polypeptides therefore include conservatively modified variants, polymorphic variants, alleles, mutants, subsequences, and interspecies homologs of TbH9 (MTB39) and Ra35 (MTB32A). The polynucleotide sequence encoding the individual polypeptides of the fusion proteins can be in any order. In some embodiments, the individual polypeptides are in order (N- to C- terminus) from large to small. Large antigens are approximately 30 to 150 kD in size, medium antigens are approximately 10 to 30 kD in size, and small antigens are approximately less than 10 kD in size. The sequence encoding the individual polypeptide may be as small as, *e.g.*, a fragment such as an individual CTL epitope encoding about 8 to 9 amino acids. The fragment may also include multiple epitopes. The fragment may also represent a larger part of the antigen sequence, *e.g.*, about 50% or more of TbH9 (MTB39) and Ra35 (MTB32A), *e.g.*, the N- and C-terminal portions of Ra35 (MTB32A).

An "Mtb72F" or "Mtb59F" fusion polypeptide of the invention specifically binds to antibodies raised against at least two antigen polypeptides, wherein each antigen polypeptide is selected from the group consisting of TbH9 (MTB39) and Ra35 (MTB32A). The antibodies can be polyclonal or monoclonal. Optionally, the fusion polypeptide

specifically binds to antibodies raised against the fusion junction of the antigens, which antibodies do not bind to the antigens individually, *i.e.*, when they are not part of a fusion protein. The fusion polypeptides optionally comprise additional polypeptides, *e.g.*, three, four, five, six, or more polypeptides, up to about 25 polypeptides, optionally heterologous polypeptides or repeated homologous polypeptides, fused to the at least two heterologous antigens. The additional polypeptides of the fusion protein are optionally derived from *Mycobacterium* as well as other sources, such as other bacterial, viral, or invertebrate, vertebrate, or mammalian sources. The individual polypeptides of the fusion protein can be in any order. As described herein, the fusion protein can also be linked to other molecules, including additional polypeptides. The compositions of the invention can also comprise additional polypeptides that are unlinked to the fusion proteins of the invention. These additional polypeptides may be heterologous or homologous polypeptides.

A polynucleotide sequence comprising a fusion protein of the invention hybridizes under stringent conditions to at least two nucleotide sequences, each encoding an antigen polypeptide selected from the group consisting of Mtb81, Mo2, TbRa3, 38 kD, Tb38-1, TbH4, HTCC#1, TbH9, MTCC#2, MTI, MSL, TbRa35, DPV, DPEP, Erd14, TbRa12, DPPD, ESAT-6, MTb82, MTb59, Mtb85 complex, and α-crystalline. The polynucleotide sequences encoding the individual antigens of the fusion polypeptide therefore include conservatively modified variants, polymorphic variants, alleles, mutants, subsequences, and interspecies homologs of Mtb81, Mo2, TbRa3, 38 kD, Tb38-1, TbH4, HTCC#1, TbH9, MTCC#2, MTI, MSL, TbRa35, DPV, DPEP, Erd14, TbRa12, DPPD, ESAT-6, MTb82, MTb59, Mtb85 complex, and α-crystalline. The polynucleotide sequence encoding the individual polypeptides of the fusion protein can be in any order. In some embodiments, the individual polypeptides are in order (N- to C- terminus) from large to small. Large antigens are approximately 30 to 150 kD in size, medium antigens are approximately 10 to 30 kD in size, and small antigens are approximately less than 10 kD in size. The sequence encoding the individual polypeptide may be as small as, *e.g.*, a fragment such as an individual CTL epitope encoding about 8 to 9 amino acids. The fragment may also include multiple epitopes. The fragment may also represent a larger part of the antigen sequence, *e.g.*, about 50% or more of Mtb81, Mo2, TbRa3, 38 kD, Tb38-1, TbH4, HTCC#1, TbH9, MTCC#2, MTI, MSL, TbRa35, DPV, DPEP, Erd14, TbRa12, DPPD, ESAT-6, MTb82, MTb59, Mtb85 complex, and α-crystalline.

A fusion polypeptide of the invention specifically binds to antibodies raised against at least two antigen polypeptides, wherein each antigen polypeptide is selected from the group consisting of Mtb81, Mo2, TbRa3, 38 kD, Tb38-1, TbH4, HTCC#1, TbH9, MTCC#2, MTI, MSL, TbRa35, DPV, DPEP, Erd14, TbRa12, DPPD, ESAT-6, MTb82, MTb59, Mtb85 complex, and α-crystalline. The antibodies can be polyclonal or monoclonal. Optionally, the fusion polypeptide specifically binds to antibodies raised against the fusion junction of the antigens, which antibodies do not bind to the antigens individually, *i.e.*, when they are not part of a fusion protein. The fusion polypeptides optionally comprise additional polypeptides, *e.g.*, three, four, five, six, or more polypeptides, up to about 25 polypeptides, optionally heterologous polypeptides or repeated homologous polypeptides, fused to the at least two heterologous antigens. The additional polypeptides of the fusion protein are optionally derived from *Mycobacterium* as well as other sources, such as other bacterial, viral, or invertebrate, vertebrate, or mammalian sources. The individual polypeptides of the fusion protein can be in any order. As described herein, the fusion protein can also be linked to other molecules, including additional polypeptides. The compositions of the invention can also comprise additional polypeptides that are unlinked to the fusion proteins of the invention. These additional polypeptides may be heterologous or homologous polypeptides.

The term "fused" refers to the covalent linkage between two polypeptides in a fusion protein. The polypeptides are typically joined via a peptide bond, either directly to each other or via an amino acid linker. Optionally, the peptides can be joined via non-peptide covalent linkages known to those of skill in the art.

"FL" refers to full-length, *i.e.*, a polypeptide that is the same length as the wild-type polypeptide.

The term "immunogenic fragment thereof" refers to a polypeptide comprising an epitope that is recognized by cytotoxic T lymphocytes, helper T lymphocytes or B cells.

The term "*Mycobacterium* species of the tuberculosis complex" includes those species traditionally considered as causing the disease tuberculosis, as well as *Mycobacterium* environmental and opportunistic species that cause tuberculosis and lung disease in immune compromised patients, such as patients with AIDS, *e.g.*, *M. tuberculosis*, *M. bovis*, or *M. africanum*, *BCG*, *M. avium*, *M. intracellulare*, *M. celatum*, *M. genavense*, *M. haemophilum*, *M. kansasii*, *M. simiae*, *M. vaccae*, *M. fortuitum*, and *M. scrofulaceum* (see, e.g., *Harrison's*

*Principles of Internal Medicine*, volume 1, pp. 1004-1014 and 1019-1023 (14th ed., Fauci *et al.*, eds., 1998).

An adjuvant refers to the components in a vaccine or therapeutic composition that increase the specific immune response to the antigen (*see, e.g.*, Edelman, *AIDS Res. Hum Retroviruses* 8:1409-1411 (1992)). Adjuvants induce immune responses of the Th1-type and Th-2 type response. Th1-type cytokines (*e.g.*, IFN-γ, IL-2, and IL-12) tend to favor the induction of cell-mediated immune response to an administered antigen, while Th-2 type cytokines (*e.g.*, IL-4, IL-5, IL-6, IL-10 and TNF-β) tend to favor the induction of humoral immune responses.

"Nucleic acid" refers to deoxyribonucleotides or ribonucleotides and polymers thereof in either single- or double-stranded form. The term encompasses nucleic acids containing known nucleotide analogs or modified backbone residues or linkages, which are synthetic, naturally occurring, and non-naturally occurring, which have similar binding properties as the reference nucleic acid, and which are metabolized in a manner similar to the reference nucleotides. Examples of such analogs include, without limitation, phosphorothioates, phosphoramidates, methyl phosphonates, chiral-methyl phosphonates, 2-O-methyl ribonucleotides, peptide-nucleic acids (PNAs).

Unless otherwise indicated, a particular nucleic acid sequence also implicitly encompasses conservatively modified variants thereof (*e.g.*, degenerate codon substitutions) and complementary sequences, as well as the sequence explicitly indicated. Specifically, degenerate codon substitutions may be achieved by generating sequences in which the third position of one or more selected (or all) codons is substituted with mixed-base and/or deoxyinosine residues (Batzer *et al.*, *Nucleic Acid Res.* 19:5081 (1991); Ohtsuka *et al.*, *J. Biol. Chem.* 260:2605-2608 (1985); Rossolini *et al.*, *Mol. Cell. Probes* 8:91-98 (1994)). The term nucleic acid is used interchangeably with gene, cDNA, mRNA, oligonucleotide, and polynucleotide.

The terms "polypeptide," "peptide" and "protein" are used interchangeably herein to refer to a polymer of amino acid residues. The terms apply to amino acid polymers in which one or more amino acid residue is an artificial chemical mimetic of a corresponding naturally occurring amino acid, as well as to naturally occurring amino acid polymers and non-naturally occurring amino acid polymer.

The term "amino acid" refers to naturally occurring and synthetic amino acids, as well as amino acid analogs and amino acid mimetics that function in a manner similar to

the naturally occurring amino acids. Naturally occurring amino acids are those encoded by the genetic code, as well as those amino acids that are later modified, *e.g.*, hydroxyproline, γ-carboxyglutamate, and O-phosphoserine. Amino acid analogs refers to compounds that have the same basic chemical structure as a naturally occurring amino acid, *i.e.*, an α carbon that is bound to a hydrogen, a carboxyl group, an amino group, and an R group, *e.g.*, homoserine, norleucine, methionine sulfoxide, methionine methyl sulfonium. Such analogs have modified R groups (*e.g.*, norleucine) or modified peptide backbones, but retain the same basic chemical structure as a naturally occurring amino acid. Amino acid mimetics refers to chemical compounds that have a structure that is different from the general chemical structure of an amino acid, but that functions in a manner similar to a naturally occurring amino acid.

Amino acids may be referred to herein by either their commonly known three letter symbols or by the one-letter symbols recommended by the IUPAC-IUB Biochemical Nomenclature Commission. Nucleotides, likewise, may be referred to by their commonly accepted single-letter codes.

"Conservatively modified variants" applies to both amino acid and nucleic acid sequences. With respect to particular nucleic acid sequences, conservatively modified variants refers to those nucleic acids which encode identical or essentially identical amino acid sequences, or where the nucleic acid does not encode an amino acid sequence, to essentially identical sequences. Because of the degeneracy of the genetic code, a large number of functionally identical nucleic acids encode any given protein. For instance, the codons GCA, GCC, GCG and GCU all encode the amino acid alanine. Thus, at every position where an alanine is specified by a codon, the codon can be altered to any of the corresponding codons described without altering the encoded polypeptide. Such nucleic acid variations are "silent variations," which are one species of conservatively modified variations. Every nucleic acid sequence herein which encodes a polypeptide also describes every possible silent variation of the nucleic acid. One of skill will recognize that each codon in a nucleic acid (except AUG, which is ordinarily the only codon for methionine, and TGG, which is ordinarily the only codon for tryptophan) can be modified to yield a functionally identical molecule. Accordingly, each silent variation of a nucleic acid which encodes a polypeptide is implicit in each described sequence.

As to amino acid sequences, one of skill will recognize that individual substitutions, deletions or additions to a nucleic acid, peptide, polypeptide, or protein sequence which alters, adds or deletes a single amino acid or a small percentage of amino acids in the encoded sequence is a "conservatively modified variant" where the alteration

results in the substitution of an amino acid with a chemically similar amino acid. Conservative substitution tables providing functionally similar amino acids are well known in the art. Such conservatively modified variants are in addition to and do not exclude polymorphic variants, interspecies homologs, and alleles of the invention.

The following eight groups each contain amino acids that are conservative substitutions for one another:

1) Alanine (A), Glycine (G);
2) Aspartic acid (D), Glutamic acid (E);
3) Asparagine (N), Glutamine (Q);
4) Arginine (R), Lysine (K);
5) Isoleucine (I), Leucine (L), Methionine (M), Valine (V);
6) Phenylalanine (F), Tyrosine (Y), Tryptophan (W);
7) Serine (S), Threonine (T); and
8) Cysteine (C), Methionine (M)

(*see, e.g.*, Creighton, *Proteins* (1984)).

The term "heterologous" when used with reference to portions of a nucleic acid indicates that the nucleic acid comprises two or more subsequences that are not found in the same relationship to each other in nature. For instance, the nucleic acid is typically recombinantly produced, having two or more sequences from unrelated genes arranged to make a new functional nucleic acid, *e.g.*, a promoter from one source and a coding region from another source. Similarly, a heterologous protein indicates that the protein comprises two or more subsequences that are not found in the same relationship to each other in nature (*e.g.*, a fusion protein).

The phrase "selectively (or specifically) hybridizes to" refers to the binding, duplexing, or hybridizing of a molecule only to a particular nucleotide sequence under stringent hybridization conditions when that sequence is present in a complex mixture (*e.g.*, total cellular or library DNA or RNA).

The phrase "stringent hybridization conditions" refers to conditions under which a probe will hybridize to its target subsequence, typically in a complex mixture of nucleic acid, but to no other sequences. Stringent conditions are sequence-dependent and will be different in different circumstances. Longer sequences hybridize specifically at higher temperatures. An extensive guide to the hybridization of nucleic acids is found in Tijssen, *Techniques in Biochemistry and Molecular Biology--Hybridization with Nucleic Probes*, "Overview of principles of hybridization and the strategy of nucleic acid assays"

(1993). Generally, stringent conditions are selected to be about 5-10°C lower than the thermal melting point ($T_m$) for the specific sequence at a defined ionic strength pH. The $T_m$ is the temperature (under defined ionic strength, pH, and nucleic concentration) at which 50% of the probes complementary to the target hybridize to the target sequence at equilibrium (as the target sequences are present in excess, at $T_m$, 50% of the probes are occupied at equilibrium). Stringent conditions will be those in which the salt concentration is less than about 1.0 M sodium ion, typically about 0.01 to 1.0 M sodium ion concentration (or other salts) at pH 7.0 to 8.3 and the temperature is at least about 30°C for short probes (*e.g.*, 10 to 50 nucleotides) and at least about 60°C for long probes (*e.g.*, greater than 50 nucleotides). Stringent conditions may also be achieved with the addition of destabilizing agents such as formamide. For selective or specific hybridization, a positive signal is at least two times background, optionally 10 times background hybridization. Exemplary stringent hybridization conditions can be as following: 50% formamide, 5x SSC, and 1% SDS, incubating at 42°C, or, 5x SSC, 1% SDS, incubating at 65°C, with wash in 0.2x SSC, and 0.1% SDS at 65°C.

Nucleic acids that do not hybridize to each other under stringent conditions are still substantially identical if the polypeptides which they encode are substantially identical. This occurs, for example, when a copy of a nucleic acid is created using the maximum codon degeneracy permitted by the genetic code. In such cases, the nucleic acids typically hybridize under moderately stringent hybridization conditions. Exemplary "moderately stringent hybridization conditions" include a hybridization in a buffer of 40% formamide, 1 M NaCl, 1% SDS at 37°C, and a wash in 1X SSC at 45°C. A positive hybridization is at least twice background. Those of ordinary skill will readily recognize that alternative hybridization and wash conditions can be utilized to provide conditions of similar stringency.

"Antibody" refers to a polypeptide comprising a framework region from an immunoglobulin gene or fragments thereof that specifically binds and recognizes an antigen. The recognized immunoglobulin genes include the kappa, lambda, alpha, gamma, delta, epsilon, and mu constant region genes, as well as the myriad immunoglobulin variable region genes. Light chains are classified as either kappa or lambda. Heavy chains are classified as gamma, mu, alpha, delta, or epsilon, which in turn define the immunoglobulin classes, IgG, IgM, IgA, IgD and IgE, respectively.

An exemplary immunoglobulin (antibody) structural unit comprises a tetramer. Each tetramer is composed of two identical pairs of polypeptide chains, each pair having one "light" (about 25 kDa) and one "heavy" chain (about 50-70 kDa). The N-

terminus of each chain defines a variable region of about 100 to 110 or more amino acids primarily responsible for antigen recognition. The terms variable light chain ($V_L$) and variable heavy chain ($V_H$) refer to these light and heavy chains respectively.

Antibodies exist, *e.g.*, as intact immunoglobulins or as a number of well-characterized fragments produced by digestion with various peptidases. Thus, for example, pepsin digests an antibody below the disulfide linkages in the hinge region to produce F(ab)'$_2$, a dimer of Fab which itself is a light chain joined to $V_H$-$C_H1$ by a disulfide bond. The F(ab)'$_2$ may be reduced under mild conditions to break the disulfide linkage in the hinge region, thereby converting the F(ab)'$_2$ dimer into an Fab' monomer. The Fab' monomer is essentially Fab with part of the hinge region (*see Fundamental Immunology* (Paul ed., 3d ed. 1993). While various antibody fragments are defined in terms of the digestion of an intact antibody, one of skill will appreciate that such fragments may be synthesized *de novo* either chemically or by using recombinant DNA methodology. Thus, the term antibody, as used herein, also includes antibody fragments either produced by the modification of whole antibodies, or those synthesized *de novo* using recombinant DNA methodologies (*e.g.*, single chain Fv) or those identified using phage display libraries (*see, e.g.*, McCafferty *et al.*, *Nature* 348:552-554 (1990)).

For preparation of monoclonal or polyclonal antibodies, any technique known in the art can be used (*see, e.g.*, Kohler & Milstein, *Nature* 256:495-497 (1975); Kozbor *et al.*, *Immunology Today* 4: 72 (1983); Cole *et al.*, pp. 77-96 in *Monoclonal Antibodies and Cancer Therapy* (1985)). Techniques for the production of single chain antibodies (U.S. Patent 4,946,778) can be adapted to produce antibodies to polypeptides of this invention. Also, transgenic mice, or other organisms such as other mammals, may be used to express humanized antibodies. Alternatively, phage display technology can be used to identify antibodies and heteromeric Fab fragments that specifically bind to selected antigens (*see, e.g.*, McCafferty *et al.*, *Nature* 348:552-554 (1990); Marks *et al.*, *Biotechnology* 10:779-783 (1992)).

The phrase "specifically (or selectively) binds" to an antibody or "specifically (or selectively) immunoreactive with," when referring to a protein or peptide, refers to a binding reaction that is determinative of the presence of the protein in a heterogeneous population of proteins and other biologics. Thus, under designated immunoassay conditions, the specified antibodies bind to a particular protein at least two times the background and do not substantially bind in a significant amount to other proteins present in the sample. Specific binding to an antibody under such conditions may require an antibody that is selected for its

specificity for a particular protein. For example, polyclonal antibodies raised to fusion proteins can be selected to obtain only those polyclonal antibodies that are specifically immunoreactive with fusion protein and not with individual components of the fusion proteins. This selection may be achieved by subtracting out antibodies that cross-react with the individual antigens. A variety of immunoassay formats may be used to select antibodies specifically immunoreactive with a particular protein. For example, solid-phase ELISA immunoassays are routinely used to select antibodies specifically immunoreactive with a protein (*see, e.g.*, Harlow & Lane, *Antibodies, A Laboratory Manual* (1988), for a description of immunoassay formats and conditions that can be used to determine specific immunoreactivity). Typically a specific or selective reaction will be at least twice background signal or noise and more typically more than 10 to 100 times background.

Polynucleotides may comprise a native sequence (*i.e.*, an endogenous sequence that encodes an individual antigen or a portion thereof) or may comprise a variant of such a sequence. Polynucleotide variants may contain one or more substitutions, additions, deletions and/or insertions such that the biological activity of the encoded fusion polypeptide is not diminished, relative to a fusion polypeptide comprising native antigens. Variants preferably exhibit at least about 70% identity, more preferably at least about 80% identity and most preferably at least about 90% identity to a polynucleotide sequence that encodes a native polypeptide or a portion thereof.

The terms "identical" or percent "identity," in the context of two or more nucleic acids or polypeptide sequences, refer to two or more sequences or subsequences that are the same or have a specified percentage of amino acid residues or nucleotides that are the same (*i.e.*, 70% identity, optionally 75%, 80%, 85%, 90%, or 95% identity over a specified region), when compared and aligned for maximum correspondence over a comparison window, or designated region as measured using one of the following sequence comparison algorithms or by manual alignment and visual inspection. Such sequences are then said to be "substantially identical." This definition also refers to the compliment of a test sequence. Optionally, the identity exists over a region that is at least about 25 to about 50 amino acids or nucleotides in length, or optionally over a region that is 75-100 amino acids or nucleotides in length.

For sequence comparison, typically one sequence acts as a reference sequence, to which test sequences are compared. When using a sequence comparison algorithm, test and reference sequences are entered into a computer, subsequence coordinates are designated, if necessary, and sequence algorithm program parameters are designated. Default program

parameters can be used, or alternative parameters can be designated. The sequence comparison algorithm then calculates the percent sequence identities for the test sequences relative to the reference sequence, based on the program parameters.

A "comparison window", as used herein, includes reference to a segment of any one of the number of contiguous positions selected from the group consisting of from 25 to 500, usually about 50 to about 200, more usually about 100 to about 150 in which a sequence may be compared to a reference sequence of the same number of contiguous positions after the two sequences are optimally aligned. Methods of alignment of sequences for comparison are well-known in the art. Optimal alignment of sequences for comparison can be conducted, *e.g.*, by the local homology algorithm of Smith & Waterman, *Adv. Appl. Math.* 2:482 (1981), by the homology alignment algorithm of Needleman & Wunsch, *J. Mol. Biol.* 48:443 (1970), by the search for similarity method of Pearson & Lipman, *Proc. Natl. Acad. Sci. USA* 85:2444 (1988), by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, WI), or by manual alignment and visual inspection (*see, e.g., Current Protocols in Molecular Biology* (Ausubel *et al.*, eds. 1995 supplement)).

One example of a useful algorithm is PILEUP. PILEUP creates a multiple sequence alignment from a group of related sequences using progressive, pairwise alignments to show relationship and percent sequence identity. It also plots a tree or dendogram showing the clustering relationships used to create the alignment. PILEUP uses a simplification of the progressive alignment method of Feng & Doolittle, *J. Mol. Evol.* 35:351-360 (1987). The method used is similar to the method described by Higgins & Sharp, *CABIOS* 5:151-153 (1989). The program can align up to 300 sequences, each of a maximum length of 5,000 nucleotides or amino acids. The multiple alignment procedure begins with the pairwise alignment of the two most similar sequences, producing a cluster of two aligned sequences. This cluster is then aligned to the next most related sequence or cluster of aligned sequences. Two clusters of sequences are aligned by a simple extension of the pairwise alignment of two individual sequences. The final alignment is achieved by a series of progressive, pairwise alignments. The program is run by designating specific sequences and their amino acid or nucleotide coordinates for regions of sequence comparison and by designating the program parameters. Using PILEUP, a reference sequence is compared to other test sequences to determine the percent sequence identity relationship using the following parameters: default gap weight (3.00), default gap length weight (0.10), and weighted end gaps. PILEUP can be

obtained from the GCG sequence analysis software package, *e.g.*, version 7.0 (Devereaux *et al.*, *Nuc. Acids Res.* 12:387-395 (1984).

Another example of algorithm that is suitable for determining percent sequence identity and sequence similarity are the BLAST and BLAST 2.0 algorithms, which are described in Altschul *et al.*, *Nuc. Acids Res.* 25:3389-3402 (1977) and Altschul *et al.*, *J. Mol. Biol.* 215:403-410 (1990), respectively. Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). This algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short words of length W in the query sequence, which either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as the neighborhood word score threshold (Altschul *et al.*, *supra*). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Cumulative scores are calculated using, for nucleotide sequences, the parameters M (reward score for a pair of matching residues; always > 0) and N (penalty score for mismatching residues; always < 0). For amino acid sequences, a scoring matrix is used to calculate the cumulative score. Extension of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm parameters W, T, and X determine the sensitivity and speed of the alignment. The BLASTN program (for nucleotide sequences) uses as defaults a wordlength (W) of 11, an expectation (E) or 10, M=5, N=-4 and a comparison of both strands. For amino acid sequences, the BLASTP program uses as defaults a wordlength of 3, and expectation (E) of 10, and the BLOSUM62 scoring matrix (*see* Henikoff & Henikoff, *Proc. Natl. Acad. Sci. USA* 89:10915 (1989)) alignments (B) of 50, expectation (E) of 10, M=5, N=-4, and a comparison of both strands.

The BLAST algorithm also performs a statistical analysis of the similarity between two sequences (*see, e.g.*, Karlin & Altschul, *Proc. Nat'l. Acad. Sci. USA* 90:5873-5787 (1993)). One measure of similarity provided by the BLAST algorithm is the smallest sum probability (P(N)), which provides an indication of the probability by which a match between two nucleotide or amino acid sequences would occur by chance. For example, a nucleic acid is considered similar to a reference sequence if the smallest sum probability in a

comparison of the test nucleic acid to the reference nucleic acid is less than about 0.2, more preferably less than about 0.01, and most preferably less than about 0.001.

## III. POLYNUCLEOTIDE COMPOSITIONS

As used herein, the terms "DNA segment" and "polynucleotide" refer to a DNA molecule that has been isolated free of total genomic DNA of a particular species. Therefore, a DNA segment encoding a polypeptide refers to a DNA segment that contains one or more coding sequences yet is substantially isolated away from, or purified free from, total genomic DNA of the species from which the DNA segment is obtained. Included within the terms "DNA segment" and "polynucleotide" are DNA segments and smaller fragments of such segments, and also recombinant vectors, including, for example, plasmids, cosmids, phagemids, phage, viruses, and the like.

As will be understood by those skilled in the art, the DNA segments of this invention can include genomic sequences, extra-genomic and plasmid-encoded sequences and smaller engineered gene segments that express, or may be adapted to express, proteins, polypeptides, peptides and the like. Such segments may be naturally isolated, or modified synthetically by the hand of man.

"Isolated," as used herein, means that a polynucleotide is substantially away from other coding sequences, and that the DNA segment does not contain large portions of unrelated coding DNA, such as large chromosomal fragments or other functional genes or polypeptide coding regions. Of course, this refers to the DNA segment as originally isolated, and does not exclude genes or coding regions later added to the segment by the hand of man.

As will be recognized by the skilled artisan, polynucleotides may be single-stranded (coding or antisense) or double-stranded, and may be DNA (genomic, cDNA or synthetic) or RNA molecules. RNA molecules include HnRNA molecules, which contain introns and correspond to a DNA molecule in a one-to-one manner, and mRNA molecules, which do not contain introns. Additional coding or non-coding sequences may, but need not, be present within a polynucleotide of the present invention, and a polynucleotide may, but need not, be linked to other molecules and/or support materials.

Polynucleotides may comprise a native sequence (*i.e.*, an endogenous sequence that encodes a *Mycobacterium* antigen or a portion thereof) or may comprise a variant, or a biological or antigenic functional equivalent of such a sequence. Polynucleotide variants may contain one or more substitutions, additions, deletions and/or insertions, as further described below, preferably such that the immunogenicity of the encoded polypeptide

is not diminished, relative to a native tumor protein. The effect on the immunogenicity of the encoded polypeptide may generally be assessed as described herein. The term "variants" also encompasses homologous genes of xenogenic origin.

In additional embodiments, the present invention provides isolated polynucleotides and polypeptides comprising various lengths of contiguous stretches of sequence identical to or complementary to one or more of the sequences disclosed herein. For example, polynucleotides are provided by this invention that comprise at least about 15, 20, 30, 40, 50, 75, 100, 150, 200, 300, 400, 500 or 1000 or more contiguous nucleotides of one or more of the sequences disclosed herein as well as all intermediate lengths there between. It will be readily understood that "intermediate lengths", in this context, means any length between the quoted values, such as 16, 17, 18, 19, etc.; 21, 22, 23, etc.; 30, 31, 32, etc.; 50, 51, 52, 53, etc.; 100, 101, 102, 103, etc.; 150, 151, 152, 153, etc.; including all integers through 200-500; 500-1,000, and the like.

The polynucleotides of the present invention, or fragments thereof, regardless of the length of the coding sequence itself, may be combined with other DNA sequences, such as promoters, polyadenylation signals, additional restriction enzyme sites, multiple cloning sites, other coding segments, and the like, such that their overall length may vary considerably. It is therefore contemplated that a nucleic acid fragment of almost any length may be employed, with the total length preferably being limited by the ease of preparation and use in the intended recombinant DNA protocol. For example, illustrative DNA segments with total lengths of about 10,000, about 5000, about 3000, about 2,000, about 1,000, about 500, about 200, about 100, about 50 base pairs in length, and the like, (including all intermediate lengths) are contemplated to be useful in many implementations of this invention.

Moreover, it will be appreciated by those of ordinary skill in the art that, as a result of the degeneracy of the genetic code, there are many nucleotide sequences that encode a polypeptide as described herein. Some of these polynucleotides bear minimal homology to the nucleotide sequence of any native gene. Nonetheless, polynucleotides that vary due to differences in codon usage are specifically contemplated by the present invention, for example polynucleotides that are optimized for human and/or primate codon selection. Further, alleles of the genes comprising the polynucleotide sequences provided herein are within the scope of the present invention. Alleles are endogenous genes that are altered as a result of one or more mutations, such as deletions, additions and/or substitutions of nucleotides. The resulting mRNA and protein may, but need not, have an altered structure or

function. Alleles may be identified using standard techniques (such as hybridization, amplification and/or database sequence comparison).

## IV. POLYNUCLEOTIDE IDENTIFICATION AND CHARACTERIZATION

Polynucleotides may be identified, prepared and/or manipulated using any of a variety of well established techniques. For example, a polynucleotide may be identified, as described in more detail below, by screening a microarray of cDNAs for tumor-associated expression (*i.e.*, expression that is at least two fold greater in a tumor than in normal tissue, as determined using a representative assay provided herein). Such screens may be performed, for example, using a Synteni microarray (Palo Alto, CA) according to the manufacturer's instructions (and essentially as described by Schena *et al.*, *Proc. Natl. Acad. Sci. USA* 93:10614-10619 (1996) and Heller *et al.*, *Proc. Natl. Acad. Sci. USA* 94:2150-2155 (1997)). Alternatively, polynucleotides may be amplified from cDNA prepared from cells expressing the proteins described herein, such as *M. tuberculosis* cells. Such polynucleotides may be amplified via polymerase chain reaction (PCR). For this approach, sequence-specific primers may be designed based on the sequences provided herein, and may be purchased or synthesized.

An amplified portion of a polynucleotide of the present invention may be used to isolate a full length gene from a suitable library (*e.g.*, a *M. tuberculosis* cDNA library) using well known techniques. Within such techniques, a library (cDNA or genomic) is screened using one or more polynucleotide probes or primers suitable for amplification. Preferably, a library is size-selected to include larger molecules. Random primed libraries may also be preferred for identifying 5' and upstream regions of genes. Genomic libraries are preferred for obtaining introns and extending 5' sequences.

For hybridization techniques, a partial sequence may be labeled (*e.g.*, by nick-translation or end-labeling with $^{32}P$) using well known techniques. A bacterial or bacteriophage library is then generally screened by hybridizing filters containing denatured bacterial colonies (or lawns containing phage plaques) with the labeled probe (*see* Sambrook *et al.*, *Molecular Cloning: A Laboratory Manual* (1989)). Hybridizing colonies or plaques are selected and expanded, and the DNA is isolated for further analysis. cDNA clones may be analyzed to determine the amount of additional sequence by, for example, PCR using a primer from the partial sequence and a primer from the vector. Restriction maps and partial sequences may be generated to identify one or more overlapping clones. The complete sequence may then be determined using standard techniques, which may involve generating a

series of deletion clones. The resulting overlapping sequences can then assembled into a single contiguous sequence. A full length cDNA molecule can be generated by ligating suitable fragments, using well known techniques.

Alternatively, there are numerous amplification techniques for obtaining a full length coding sequence from a partial cDNA sequence. Within such techniques, amplification is generally performed via PCR. Any of a variety of commercially available kits may be used to perform the amplification step. Primers may be designed using, for example, software well known in the art. Primers are preferably 22-30 nucleotides in length, have a GC content of at least 50% and anneal to the target sequence at temperatures of about 68°C to 72°C. The amplified region may be sequenced as described above, and overlapping sequences assembled into a contiguous sequence.

One such amplification technique is inverse PCR (*see* Triglia *et al.*, *Nucl. Acids Res.* 16:8186 (1988)), which uses restriction enzymes to generate a fragment in the known region of the gene. The fragment is then circularized by intramolecular ligation and used as a template for PCR with divergent primers derived from the known region. Within an alternative approach, sequences adjacent to a partial sequence may be retrieved by amplification with a primer to a linker sequence and a primer specific to a known region. The amplified sequences are typically subjected to a second round of amplification with the same linker primer and a second primer specific to the known region. A variation on this procedure, which employs two primers that initiate extension in opposite directions from the known sequence, is described in WO 96/38591. Another such technique is known as "rapid amplification of cDNA ends" or RACE. This technique involves the use of an internal primer and an external primer, which hybridizes to a polyA region or vector sequence, to identify sequences that are 5' and 3' of a known sequence. Additional techniques include capture PCR (Lagerstrom *et al.*, *PCR Methods Applic.* 1:111-19 (1991)) and walking PCR (Parker *et al.*, *Nucl. Acids. Res.* 19:3055-60 (1991)). Other methods employing amplification may also be employed to obtain a full length cDNA sequence.

In certain instances, it is possible to obtain a full length cDNA sequence by analysis of sequences provided in an expressed sequence tag (EST) database, such as that available from GenBank. Searches for overlapping ESTs may generally be performed using well known programs (*e.g.*, NCBI BLAST searches), and such ESTs may be used to generate a contiguous full length sequence. Full length DNA sequences may also be obtained by analysis of genomic fragments.

## V. POLYNUCLEOTIDE EXPRESSION IN HOST CELLS

In other embodiments of the invention, polynucleotide sequences or fragments thereof which encode polypeptides of the invention, or fusion proteins or functional equivalents thereof, may be used in recombinant DNA molecules to direct expression of a polypeptide in appropriate host cells. Due to the inherent degeneracy of the genetic code, other DNA sequences that encode substantially the same or a functionally equivalent amino acid sequence may be produced and these sequences may be used to clone and express a given polypeptide.

As will be understood by those of skill in the art, it may be advantageous in some instances to produce polypeptide-encoding nucleotide sequences possessing non-naturally occurring codons. For example, codons preferred by a particular prokaryotic or eukaryotic host can be selected to increase the rate of protein expression or to produce a recombinant RNA transcript having desirable properties, such as a half-life which is longer than that of a transcript generated from the naturally occurring sequence.

Moreover, the polynucleotide sequences of the present invention can be engineered using methods generally known in the art in order to alter polypeptide encoding sequences for a variety of reasons, including but not limited to, alterations which modify the cloning, processing, and/or expression of the gene product. For example, DNA shuffling by random fragmentation and PCR reassembly of gene fragments and synthetic oligonucleotides may be used to engineer the nucleotide sequences. In addition, site-directed mutagenesis may be used to insert new restriction sites, alter glycosylation patterns, change codon preference, produce splice variants, or introduce mutations, and so forth.

In another embodiment of the invention, natural, modified, or recombinant nucleic acid sequences may be ligated to a heterologous sequence to encode a fusion protein. For example, to screen peptide libraries for inhibitors of polypeptide activity, it may be useful to encode a chimeric protein that can be recognized by a commercially available antibody. A fusion protein may also be engineered to contain a cleavage site located between the polypeptide-encoding sequence and the heterologous protein sequence, so that the polypeptide may be cleaved and purified away from the heterologous moiety.

Sequences encoding a desired polypeptide may be synthesized, in whole or in part, using chemical methods well known in the art (*see* Caruthers, M. H. *et al.*, *Nucl. Acids Res. Symp. Ser.* pp. 215-223 (1980), Horn *et al.*, *Nucl. Acids Res. Symp. Ser.* pp. 225-232 (1980)). Alternatively, the protein itself may be produced using chemical methods to

synthesize the amino acid sequence of a polypeptide, or a portion thereof. For example, peptide synthesis can be performed using various solid-phase techniques (Roberge *et al., Science* 269:202-204 (1995)) and automated synthesis may be achieved, for example, using the ABI 431A Peptide Synthesizer (Perkin Elmer, Palo Alto, CA).

A newly synthesized peptide may be substantially purified by preparative high performance liquid chromatography (*e.g.*, Creighton, *Proteins. Structures and Molecular Principles* (1983)) or other comparable techniques available in the art. The composition of the synthetic peptides may be confirmed by amino acid analysis or sequencing (*e.g.*, the Edman degradation procedure). Additionally, the amino acid sequence of a polypeptide, or any part thereof, may be altered during direct synthesis and/or combined using chemical methods with sequences from other proteins, or any part thereof, to produce a variant polypeptide.

In order to express a desired polypeptide, the nucleotide sequences encoding the polypeptide, or functional equivalents, may be inserted into appropriate expression vector, *i.e.*, a vector which contains the necessary elements for the transcription and translation of the inserted coding sequence. Methods which are well known to those skilled in the art may be used to construct expression vectors containing sequences encoding a polypeptide of interest and appropriate transcriptional and translational control elements. These methods include in vitro recombinant DNA techniques, synthetic techniques, and in vivo genetic recombination. Such techniques are described in Sambrook *et al., Molecular Cloning, A Laboratory Manual* (1989), and Ausubel *et al., Current Protocols in Molecular Biology* (1989).

A variety of expression vector/host systems may be utilized to contain and express polynucleotide sequences. These include, but are not limited to, microorganisms such as bacteria transformed with recombinant bacteriophage, plasmid, or cosmid DNA expression vectors; yeast transformed with yeast expression vectors; insect cell systems infected with virus expression vectors (*e.g.*, baculovirus); plant cell systems transformed with virus expression vectors (*e.g.*, cauliflower mosaic virus, CaMV; tobacco mosaic virus, TMV) or with bacterial expression vectors (*e.g.*, Ti or pBR322 plasmids); or animal cell systems.

The "control elements" or "regulatory sequences" present in an expression vector are those non-translated regions of the vector--enhancers, promoters, 5' and 3' untranslated regions--which interact with host cellular proteins to carry out transcription and translation. Such elements may vary in their strength and specificity. Depending on the vector system and host utilized, any number of suitable transcription and translation elements, including constitutive and inducible promoters, may be used. For example, when cloning in

bacterial systems, inducible promoters such as the hybrid lacZ promoter of the PBLUESCRIPT phagemid (Stratagene, La Jolla, Calif.) or PSPORT1 plasmid (Gibco BRL, Gaithersburg, MD) and the like may be used. In mammalian cell systems, promoters from mammalian genes or from mammalian viruses are generally preferred. If it is necessary to generate a cell line that contains multiple copies of the sequence encoding a polypeptide, vectors based on SV40 or EBV may be advantageously used with an appropriate selectable marker.

In bacterial systems, a number of expression vectors may be selected depending upon the use intended for the expressed polypeptide. For example, when large quantities are needed, for example for the induction of antibodies, vectors which direct high level expression of fusion proteins that are readily purified may be used. Such vectors include, but are not limited to, the multifunctional *E. coli* cloning and expression vectors such as BLUESCRIPT (Stratagene), in which the sequence encoding the polypeptide of interest may be ligated into the vector in frame with sequences for the amino-terminal Met and the subsequent 7 residues of β-galactosidase so that a hybrid protein is produced; pIN vectors (Van Heeke &Schuster, *J. Biol. Chem.* 264:5503-5509 (1989)); and the like. pGEX Vectors (Promega, Madison, Wis.) may also be used to express foreign polypeptides as fusion proteins with glutathione S-transferase (GST). In general, such fusion proteins are soluble and can easily be purified from lysed cells by adsorption to glutathione-agarose beads followed by elution in the presence of free glutathione. Proteins made in such systems may be designed to include heparin, thrombin, or factor XA protease cleavage sites so that the cloned polypeptide of interest can be released from the GST moiety at will.

In the yeast, *Saccharomyces cerevisiae*, a number of vectors containing constitutive or inducible promoters such as alpha factor, alcohol oxidase, and PGH may be used. For reviews, see Ausubel *et al.* (*supra*) and Grant *et al.*, *Methods Enzymol.* 153:516-544 (1987).

In cases where plant expression vectors are used, the expression of sequences encoding polypeptides may be driven by any of a number of promoters. For example, viral promoters such as the 35S and 19S promoters of CaMV may be used alone or in combination with the omega leader sequence from TMV (Takamatsu, *EMBO J.* 6:307-311 (1987)). Alternatively, plant promoters such as the small subunit of RUBISCO or heat shock promoters may be used (Coruzzi *et al.*, *EMBO J.* 3:1671-1680 (1984); Broglie *et al.*, *Science* 224:838-843 (1984); and Winter *et al.*, *Results Probl. Cell Differ.* 17:85-105 (1991)). These

constructs can be introduced into plant cells by direct DNA transformation or pathogen-mediated transfection. Such techniques are described in a number of generally available reviews (*see, e.g.*, Hobbs *in McGraw Hill Yearbook of Science and Technology* pp. 191-196 (1992)).

An insect system may also be used to express a polypeptide of interest. For example, in one such system, *Autographa californica* nuclear polyhedrosis virus (AcNPV) is used as a vector to express foreign genes in *Spodoptera frugiperda* cells or in *Trichoplusia larvae*. The sequences encoding the polypeptide may be cloned into a non-essential region of the virus, such as the polyhedrin gene, and placed under control of the polyhedrin promoter. Successful insertion of the polypeptide-encoding sequence will render the polyhedrin gene inactive and produce recombinant virus lacking coat protein. The recombinant viruses may then be used to infect, for example, *S. frugiperda* cells or *Trichoplusia larvae* in which the polypeptide of interest may be expressed (Engelhard *et al., Proc. Natl. Acad. Sci. U.S.A.* 91 :3224-3227 (1994)).

In mammalian host cells, a number of viral-based expression systems are generally available. For example, in cases where an adenovirus is used as an expression vector, sequences encoding a polypeptide of interest may be ligated into an adenovirus transcription/translation complex consisting of the late promoter and tripartite leader sequence. Insertion in a non-essential E1 or E3 region of the viral genome may be used to obtain a viable virus which is capable of expressing the polypeptide in infected host cells (Logan & Shenk, *Proc. Natl. Acad. Sci. U.S.A.* 81:3655-3659 (1984)). In addition, transcription enhancers, such as the Rous sarcoma virus (RSV) enhancer, may be used to increase expression in mammalian host cells.

Specific initiation signals may also be used to achieve more efficient translation of sequences encoding a polypeptide of interest. Such signals include the ATG initiation codon and adjacent sequences. In cases where sequences encoding the polypeptide, its initiation codon, and upstream sequences are inserted into the appropriate expression vector, no additional transcriptional or translational control signals may be needed. However, in cases where only coding sequence, or a portion thereof, is inserted, exogenous translational control signals including the ATG initiation codon should be provided. Furthermore, the initiation codon should be in the correct reading frame to ensure translation of the entire insert. Exogenous translational elements and initiation codons may be of various origins, both natural and synthetic. The efficiency of expression may be enhanced by the inclusion of

enhancers which are appropriate for the particular cell system which is used, such as those described in the literature (Scharf, *et al., Results Probl. Cell Differ.* 20:125-162 (1994)).

In addition, a host cell strain may be chosen for its ability to modulate the expression of the inserted sequences or to process the expressed protein in the desired fashion. Such modifications of the polypeptide include, but are not limited to, acetylation, carboxylation, glycosylation, phosphorylation, lipidation, and acylation. Post-translational processing which cleaves a "prepro" form of the protein may also be used to facilitate correct insertion, folding and/or function. Different host cells such as CHO, HeLa, MDCK, HEK293, and WI38, which have specific cellular machinery and characteristic mechanisms for such post-translational activities, may be chosen to ensure the correct modification and processing of the foreign protein.

For long-term, high-yield production of recombinant proteins, stable expression is generally preferred. For example, cell lines which stably express a polynucleotide of interest may be transformed using expression vectors which may contain viral origins of replication and/or endogenous expression elements and a selectable marker gene on the same or on a separate vector. Following the introduction of the vector, cells may be allowed to grow for 1-2 days in an enriched media before they are switched to selective media. The purpose of the selectable marker is to confer resistance to selection, and its presence allows growth and recovery of cells which successfully express the introduced sequences. Resistant clones of stably transformed cells may be proliferated using tissue culture techniques appropriate to the cell type.

Any number of selection systems may be used to recover transformed cell lines. These include, but are not limited to, the herpes simplex virus thymidine kinase (Wigler *et al., Cell* 11:223-32 (1977)) and adenine phosphoribosyltransferase (Lowy *et al., Cell* 22:817-23 (1990)) genes which can be employed in tk.sup.- or aprt.sup.- cells, respectively. Also, antimetabolite, antibiotic or herbicide resistance can be used as the basis for selection; for example, dhfr which confers resistance to methotrexate (Wigler *et al., Proc. Natl. Acad. Sci. U.S.A.* 77:3567-70 (1980)); npt, which confers resistance to the aminoglycosides, neomycin and G-418 (Colbere-Garapin *et al., J. Mol. Biol.* 150:1-14 (1981)); and als or pat, which confer resistance to chlorsulfuron and phosphinotricin acetyltransferase, respectively (Murry, *supra*). Additional selectable genes have been described, for example, trpB, which allows cells to utilize indole in place of tryptophan, or hisD, which allows cells to utilize histinol in place of histidine (Hartman & Mulligan, *Proc. Natl. Acad. Sci. U.S.A.* 85:8047-51 (1988)). Recently, the use of visible markers has gained popularity with such markers as

anthocyanins, β-glucuronidase and its substrate GUS, and luciferase and its substrate luciferin, being widely used not only to identify transformants, but also to quantify the amount of transient or stable protein expression attributable to a specific vector system (Rhodes *et al., Methods Mol. Biol.* 55:121-131 (1995)).

Although the presence/absence of marker gene expression suggests that the gene of interest is also present, its presence and expression may need to be confirmed. For example, if the sequence encoding a polypeptide is inserted within a marker gene sequence, recombinant cells containing sequences can be identified by the absence of marker gene function. Alternatively, a marker gene can be placed in tandem with a polypeptide-encoding sequence under the control of a single promoter. Expression of the marker gene in response to induction or selection usually indicates expression of the tandem gene as well.

Alternatively, host cells which contain and express a desired polynucleotide sequence may be identified by a variety of procedures known to those of skill in the art. These procedures include, but are not limited to, DNA-DNA or DNA-RNA hybridizations and protein bioassay or immunoassay techniques which include membrane, solution, or chip based technologies for the detection and/or quantification of nucleic acid or protein.

A variety of protocols for detecting and measuring the expression of polynucleotide-encoded products, using either polyclonal or monoclonal antibodies specific for the product are known in the art. Examples include enzyme-linked immunosorbent assay (ELISA), radioimmunoassay (RIA), and fluorescence activated cell sorting (FACS). A two-site, monoclonal-based immunoassay utilizing monoclonal antibodies reactive to two non-interfering epitopes on a given polypeptide may be preferred for some applications, but a competitive binding assay may also be employed. These and other assays are described, among other places, in Hampton *et al., Serological Methods, a Laboratory Manual* (1990) and Maddox *et al., J. Exp. Med.* 158:1211-1216 (1983).

A wide variety of labels and conjugation techniques are known by those skilled in the art and may be used in various nucleic acid and amino acid assays. Means for producing labeled hybridization or PCR probes for detecting sequences related to polynucleotides include oligolabeling, nick translation, end-labeling or PCR amplification using a labeled nucleotide. Alternatively, the sequences, or any portions thereof may be cloned into a vector for the production of an mRNA probe. Such vectors are known in the art, are commercially available, and may be used to synthesize RNA probes in vitro by addition of an appropriate RNA polymerase such as T7, T3, or SP6 and labeled nucleotides.

These procedures may be conducted using a variety of commercially available kits. Suitable reporter molecules or labels, which may be used include radionuclides, enzymes, fluorescent, chemiluminescent, or chromogenic agents as well as substrates, cofactors, inhibitors, magnetic particles, and the like.

Host cells transformed with a polynucleotide sequence of interest may be cultured under conditions suitable for the expression and recovery of the protein from cell culture. The protein produced by a recombinant cell may be secreted or contained intracellularly depending on the sequence and/or the vector used. As will be understood by those of skill in the art, expression vectors containing polynucleotides of the invention may be designed to contain signal sequences which direct secretion of the encoded polypeptide through a prokaryotic or eukaryotic cell membrane. Other recombinant constructions may be used to join sequences encoding a polypeptide of interest to nucleotide sequence encoding a polypeptide domain which will facilitate purification of soluble proteins. Such purification facilitating domains include, but are not limited to, metal chelating peptides such as histidine-tryptophan modules that allow purification on immobilized metals, protein A domains that allow purification on immobilized immunoglobulin, and the domain utilized in the FLAGS extension/affinity purification system (Immunex Corp., Seattle, Washington). The inclusion of cleavable linker sequences such as those specific for Factor XA or enterokinase (Invitrogen. San Diego, Calif.) between the purification domain and the encoded polypeptide may be used to facilitate purification. One such expression vector provides for expression of a fusion protein containing a polypeptide of interest and a nucleic acid encoding 6 histidine residues preceding a thioredoxin or an enterokinase cleavage site. The histidine residues facilitate purification on IMIAC (immobilized metal ion affinity chromatography) as described in Porath *et al., Prot. Exp. Purif.* 3:263-281 (1992) while the enterokinase cleavage site provides a means for purifying the desired polypeptide from the fusion protein. A discussion of vectors which contain fusion proteins is provided in Kroll *et al., DNA Cell Biol.* 12:441-453 (1993)).

In addition to recombinant production methods, polypeptides of the invention, and fragments thereof, may be produced by direct peptide synthesis using solid-phase techniques (Merrifield, *J. Am. Chem. Soc.* 85:2149-2154 (1963)). Protein synthesis may be performed using manual techniques or by automation. Automated synthesis may be achieved, for example, using Applied Biosystems 431A Peptide Synthesizer (Perkin Elmer). Alternatively, various fragments may be chemically synthesized separately and combined using chemical methods to produce the full length molecule.

## VI. IN VIVO POLYNUCLEOTIDE DELIVERY TECHNIQUES

In additional embodiments, genetic constructs comprising one or more of the polynucleotides of the invention are introduced into cells *in vivo*. This may be achieved using any of a variety or well known approaches, several of which are outlined below for the purpose of illustration.

### 1. Adenovirus

One of the preferred methods for *in vivo* delivery of one or more nucleic acid sequences involves the use of an adenovirus expression vector. "Adenovirus expression vector" is meant to include those constructs containing adenovirus sequences sufficient to (a) support packaging of the construct and (b) to express a polynucleotide that has been cloned therein in a sense or antisense orientation. Of course, in the context of an antisense construct, expression does not require that the gene product be synthesized.

The expression vector comprises a genetically engineered form of an adenovirus. Knowledge of the genetic organization of adenovirus, a 36 kb, linear, double-stranded DNA virus, allows substitution of large pieces of adenoviral DNA with foreign sequences up to 7 kb (Grunhaus & Horwitz, 1992). In contrast to retrovirus, the adenoviral infection of host cells does not result in chromosomal integration because adenoviral DNA can replicate in an episomal manner without potential genotoxicity. Also, adenoviruses are structurally stable, and no genome rearrangement has been detected after extensive amplification. Adenovirus can infect virtually all epithelial cells regardless of their cell cycle stage. So far, adenoviral infection appears to be linked only to mild disease such as acute respiratory disease in humans.

Adenovirus is particularly suitable for use as a gene transfer vector because of its mid-sized genome, ease of manipulation, high titer, wide target-cell range and high infectivity. Both ends of the viral genome contain 100-200 base pair inverted repeats (ITRs), which are *cis* elements necessary for viral DNA replication and packaging. The early (E) and late (L) regions of the genome contain different transcription units that are divided by the onset of viral DNA replication. The E1 region (E1A and E1B) encodes proteins responsible for the regulation of transcription of the viral genome and a few cellular genes. The expression of the E2 region (E2A and E2B) results in the synthesis of the proteins for viral DNA replication. These proteins are involved in DNA replication, late gene expression and host cell shut-off (Renan, 1990). The products of the late genes, including the majority of the viral capsid proteins, are expressed only after significant processing of a single primary

transcript issued by the major late promoter (MLP). The MLP, (located at 16.8 m.u.) is particularly efficient during the late phase of infection, and all the mRNA's issued from this promoter possess a 5'-tripartite leader (TPL) sequence which makes them preferred mRNA's for translation.

In a current system, recombinant adenovirus is generated from homologous recombination between shuttle vector and provirus vector. Due to the possible recombination between two proviral vectors, wild-type adenovirus may be generated from this process. Therefore, it is critical to isolate a single clone of virus from an individual plaque and examine its genomic structure.

Generation and propagation of the current adenovirus vectors, which are replication deficient, depend on a unique helper cell line, designated 293, which was transformed from human embryonic kidney cells by Ad5 DNA fragments and constitutively expresses E1 proteins (Graham *et al.*, 1977). Since the E3 region is dispensable from the adenovirus genome (Jones & Shenk, 1978), the current adenovirus vectors, with the help of 293 cells, carry foreign DNA in either the E1, the D3 or both regions (Graham & Prevec, 1991). In nature, adenovirus can package approximately 105% of the wild-type genome (Ghosh-Choudhury *et al.*, 1987), providing capacity for about 2 extra kB of DNA. Combined with the approximately 5.5 kB of DNA that is replaceable in the E1 and E3 regions, the maximum capacity of the current adenovirus vector is under 7.5 kB, or about 15% of the total length of the vector. More than 80% of the adenovirus viral genome remains in the vector backbone and is the source of vector-borne cytotoxicity. Also, the replication deficiency of the E1-deleted virus is incomplete. For example, leakage of viral gene expression has been observed with the currently available vectors at high multiplicities of infection (MOI) (Mulligan, 1993).

Helper cell lines may be derived from human cells such as human embryonic kidney cells, muscle cells, hematopoietic cells or other human embryonic mesenchymal or epithelial cells. Alternatively, the helper cells may be derived from the cells of other mammalian species that are permissive for human adenovirus. Such cells include, *e.g.*, Vero cells or other monkey embryonic mesenchymal or epithelial cells. As stated above, the currently preferred helper cell line is 293.

Recently, Racher *et al.* (1995) disclosed improved methods for culturing 293 cells and propagating adenovirus. In one format, natural cell aggregates are grown by inoculating individual cells into 1 liter siliconized spinner flasks (Techne, Cambridge, UK)

containing 100-200 ml of medium. Following stirring at 40 rpm, the cell viability is estimated with trypan blue. In another format, Fibra-Cel microcarriers (Bibby Sterlin, Stone, UK) (5 g/l) is employed as follows. A cell inoculum, resuspended in 5 ml of medium, is added to the carrier (50 ml) in a 250 ml Erlenmeyer flask and left stationary, with occasional agitation, for 1 to 4 h. The medium is then replaced with 50 ml of fresh medium and shaking initiated. For virus production, cells are allowed to grow to about 80% confluence, after which time the medium is replaced (to 25% of the final volume) and adenovirus added at an MOI of 0.05. Cultures are left stationary overnight, following which the volume is increased to 100% and shaking commenced for another 72 h.

Other than the requirement that the adenovirus vector be replication defective, or at least conditionally defective, the nature of the adenovirus vector is not believed to be crucial to the successful practice of the invention. The adenovirus may be of any of the 42 different known serotypes or subgroups A-F. Adenovirus type 5 of subgroup C is the preferred starting material in order to obtain a conditional replication-defective adenovirus vector for use in the present invention, since Adenovirus type 5 is a human adenovirus about which a great deal of biochemical and genetic information is known, and it has historically been used for most constructions employing adenovirus as a vector.

As stated above, the typical vector according to the present invention is replication defective and will not have an adenovirus E1 region. Thus, it will be most convenient to introduce the polynucleotide encoding the gene of interest at the position from which the E1-coding sequences have been removed. However, the position of insertion of the construct within the adenovirus sequences is not critical to the invention. The polynucleotide encoding the gene of interest may also be inserted in lieu of the deleted E3 region in E3 replacement vectors as described by Karlsson *et al.* (1986) or in the E4 region where a helper cell line or helper virus complements the E4 defect.

Adenovirus is easy to grow and manipulate and exhibits broad host range *in vitro* and *in vivo*. This group of viruses can be obtained in high titers, *e.g.*, $10^9$-$10^{11}$ plaque-forming units per ml, and they are highly infective. The life cycle of adenovirus does not require integration into the host cell genome. The foreign genes delivered by adenovirus vectors are episomal and, therefore, have low genotoxicity to host cells. No side effects have been reported in studies of vaccination with wild-type adenovirus (Couch *et al.*, 1963; Top *et al.*, 1971), demonstrating their safety and therapeutic potential as *in vivo* gene transfer vectors.

Adenovirus vectors have been used in eukaryotic gene expression (Levrero *et al.*, 1991; Gomez-Foix *et al.*, 1992) and vaccine development (Grunhaus & Horwitz, 1992; Graham & Prevec, 1992). Recently, animal studies suggested that recombinant adenovirus could be used for gene therapy (Stratford-Perricaudet & Perricaudet, 1991; Stratford-Perricaudet *et al.*, 1990; Rich *et al.*, 1993). Studies in administering recombinant adenovirus to different tissues include trachea instillation (Rosenfeld *et al.*, 1991; Rosenfeld *et al.*, 1992), muscle injection (Ragot *et al.*, 1993), peripheral intravenous injections (Herz & Gerard, 1993) and stereotactic inoculation into the brain (Le Gal La Salle *et al.*, 1993).

**B. Retroviruses**

The retroviruses are a group of single-stranded RNA viruses characterized by an ability to convert their RNA to double-stranded DNA in infected cells by a process of reverse-transcription (Coffin, 1990). The resulting DNA then stably integrates into cellular chromosomes as a provirus and directs synthesis of viral proteins. The integration results in the retention of the viral gene sequences in the recipient cell and its descendants. The retroviral genome contains three genes, gag, pol, and env that code for capsid proteins, polymerase enzyme, and envelope components, respectively. A sequence found upstream from the gag gene contains a signal for packaging of the genome into virions. Two long terminal repeat (LTR) sequences are present at the 5' and 3' ends of the viral genome. These contain strong promoter and enhancer sequences and are also required for integration in the host cell genome (Coffin, 1990).

In order to construct a retroviral vector, a nucleic acid encoding one or more oligonucleotide or polynucleotide sequences of interest is inserted into the viral genome in the place of certain viral sequences to produce a virus that is replication-defective. In order to produce virions, a packaging cell line containing the gag, pol, and env genes but without the LTR and packaging components is constructed (Mann *et al.*, 1983). When a recombinant plasmid containing a cDNA, together with the retroviral LTR and packaging sequences is introduced into this cell line (by calcium phosphate precipitation for example), the packaging sequence allows the RNA transcript of the recombinant plasmid to be packaged into viral particles, which are then secreted into the culture media (Nicolas & Rubenstein, 1988; Temin, 1986; Mann *et al.*, 1983). The media containing the recombinant retroviruses is then collected, optionally concentrated, and used for gene transfer. Retroviral vectors are able to infect a broad variety of cell types. However, integration and stable expression require the division of host cells (Paskind *et al.*, 1975).

A novel approach designed to allow specific targeting of retrovirus vectors was recently developed based on the chemical modification of a retrovirus by the chemical addition of lactose residues to the viral envelope. This modification could permit the specific infection of hepatocytes *via* sialoglycoprotein receptors.

A different approach to targeting of recombinant retroviruses was designed in which biotinylated antibodies against a retroviral envelope protein and against a specific cell receptor were used. The antibodies were coupled *via* the biotin components by using streptavidin (Roux *et al.*, 1989). Using antibodies against major histocompatibility complex class I and class II antigens, they demonstrated the infection of a variety of human cells that bore those surface antigens with an ecotropic virus *in vitro* (Roux *et al.*, 1989).

### C. Adeno-Associated Viruses

AAV (Ridgeway, 1988; Hermonat & Muzyczka, 1984) is a parovirus, discovered as a contamination of adenoviral stocks. It is a ubiquitous virus (antibodies are present in 85% of the US human population) that has not been linked to any disease. It is also classified as a dependovirus, because its replications is dependent on the presence of a helper virus, such as adenovirus. Five serotypes have been isolated, of which AAV-2 is the best characterized. AAV has a single-stranded linear DNA that is encapsidated into capsid proteins VP1, VP2 and VP3 to form an icosahedral virion of 20 to 24 nm in diameter (Muzyczka & McLaughlin, 1988).

The AAV DNA is approximately 4.7 kilobases long. It contains two open reading frames and is flanked by two ITRs. There are two major genes in the AAV genome: *rep* and *cap*. The *rep* gene codes for proteins responsible for viral replications, whereas *cap* codes for capsid protein VP1-3. Each ITR forms a T-shaped hairpin structure. These terminal repeats are the only essential *cis* components of the AAV for chromosomal integration. Therefore, the AAV can be used as a vector with all viral coding sequences removed and replaced by the cassette of genes for delivery. Three viral promoters have been identified and named p5, p19, and p40, according to their map position. Transcription from p5 and p19 results in production of rep proteins, and transcription from p40 produces the capsid proteins (Hermonat & Muzyczka, 1984).

There are several factors that prompted researchers to study the possibility of using rAAV as an expression vector. One is that the requirements for delivering a gene to integrate into the host chromosome are surprisingly few. It is necessary to have the 145-bp ITRs, which are only 6% of the AAV genome. This leaves room in the vector to assemble a